Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders

Emile Mathieu†, Charline Le Lan†, Chris J. Maddison†*, Ryota Tomioka‡ and Yee Whye Teh†*

† Department of Statistics, University of Oxford, United Kingdom, * DeepMind, London, United Kingdom, ‡ Microsoft Research, Cambridge, United Kingdom

UNIVERSITY OF OXFORD

Overview & Contributions

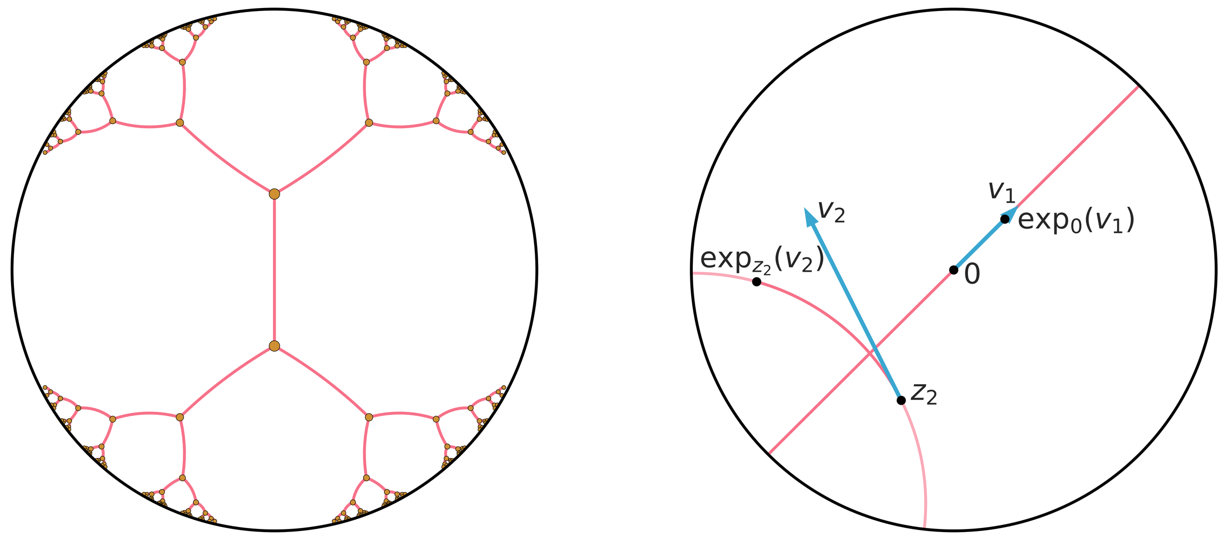
Many real datasets are **hierarchically structured**. However, traditional variational auto-encoders (VAEs) [1, 2] map data in a **Euclidean latent space** which cannot efficiently embed tree-like structures. **Hyperbolic spaces** with negative curvature can [3] and their smoothness is well-suited for gradient based approaches [4].

1. We empirically demonstrate that endowing a VAE with a **Poincaré ball latent space** can be beneficial in terms of model generalisation and can yield more interpretable representations.
2. We propose **efficient and reparametrisable sampling schemes**, and calculate the **probability density functions**, for two canonical Gaussian generalisations defined on the Poincaré ball, namely the **maximum-entropy** and **wrapped** normal distributions.
3. We introduce a **decoder architecture** taking into account the hyperbolic geometry, which we empirically show to be crucial.

The Poincaré ball model of hyperbolic geometry

The d-dimensional Poincaré ball with curvature $-c$ is the Riemannian manifold $\mathbb{B}_c^d = (\mathcal{B}_c^d, \mathfrak{g}_p^c)$ [5], where $\mathcal{B}_c^d = \{\boldsymbol{z} \in \mathbb{R}^d \mid \|\boldsymbol{z}\|_2 \leq 1/\sqrt{c}\}$, and $\mathfrak{g}_p^c$ its *metric tensor*,

$$\mathfrak{g}_p^c(\boldsymbol{z}) = (\lambda_{\boldsymbol{z}}^c)^2 \, \mathfrak{g}_e(\boldsymbol{z}) = \left(\frac{2}{1 - c\,\|\boldsymbol{z}\|^2}\right)^2 \mathfrak{g}_e(\boldsymbol{z}) \qquad (1)$$



(a) Isometric embedding of tree. (b) Geodesics and exponential maps.

We are extremely grateful to Adam Foster, Phillipe Gagnon and Emmanuel Chevallier for their help. EM, YWT's research leading to these results received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007- 2013) ERC grant agreement no. 617071 and they acknowledge Microsoft Research and EPSRC for funding EM's studentship, and EPSRC grant agreement no. EP/N509711/1 for funding CL's studentship.

[1] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.
[2] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *ICML*, 2014.
[3] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In Marc van Kreveld and Bettina Speckmann, editors, *Graph Drawing*, pages 355–366. Springer Berlin Heidelberg, 2012.
[4] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, pages 6341–6350. 2017.
[5] E. Beltrami. *Teoria fondamentale degli spazii di curvatura costante: memoria*. F. Zanetti, 1868.
[6] Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127, Jul 2006.
[7] Salem Said, Lionel Bombrun, and Yannick Berthoumieu. New riemannian priors on the univariate normal model. *Entropy*, 16(7):4015–4031, 2014.
[8] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *International Conference on Neural Information Processing Systems*, pages 5350–5360, 2018.
[9] Michael Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. In *International Conference on Neural Information Processing Systems*, pages 439–450, 2018.
[10] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *Workshop on Bayesian Deep Learning, NIPS*, 2016.

Hyperbolic normal distributions reparametrisation

Invariant measure: $d\mathcal{M}(\boldsymbol{z}) = \sqrt{|G(\boldsymbol{z})|}d\boldsymbol{z} = (\lambda_{\boldsymbol{z}}^c)^d d\boldsymbol{z}$ with $d\boldsymbol{z}$ the Lebesgue measure

Riemannian normal: $\mathcal{N}^{\mathsf{R}}_{\mathbb{B}_c^d}(\boldsymbol{z}|\boldsymbol{\mu}, \sigma^2) \propto \exp\left(-d_p^c(\boldsymbol{\mu}, \boldsymbol{z})^2/2\sigma^2\right)$ (maximum-entropy [6, 7])

Wrapped normal: $\boldsymbol{z} = \exp_{\boldsymbol{\mu}}^c\left(\boldsymbol{v}/\lambda_{\boldsymbol{\mu}}^c\right)$ with $\boldsymbol{v} \sim \mathcal{N}(\cdot|\boldsymbol{0}, \Sigma)$ (push-forward)

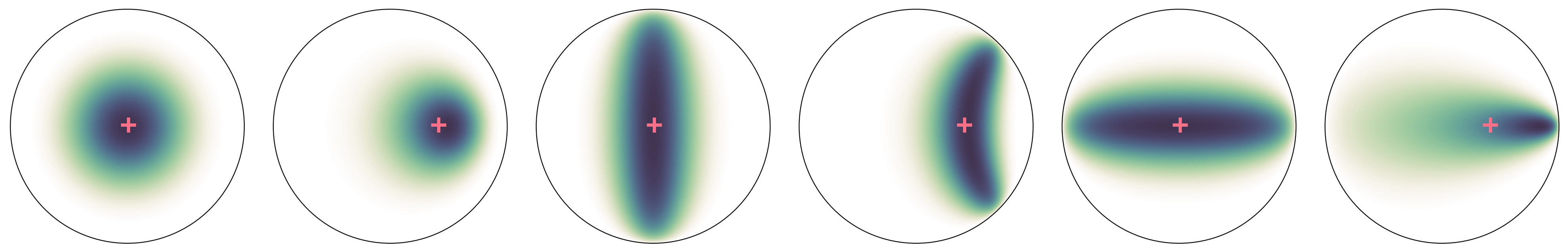
$$\mathcal{N}^{\mathsf{W}}_{\mathbb{B}_c^d}(\boldsymbol{z}|\boldsymbol{\mu}, \Sigma) = \mathcal{N}\left(\lambda_{\boldsymbol{\mu}}^c \log_{\boldsymbol{\mu}}(\boldsymbol{z})\middle|\boldsymbol{0}, \Sigma\right)\left(\sqrt{c}\, d_p^c(\boldsymbol{\mu}, \boldsymbol{z})/\sinh(\sqrt{c}\, d_p^c(\boldsymbol{\mu}, \boldsymbol{z}))\right)^{d-1}$$

Reparametrisation through the exponential map: $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{B}_c^d}(\boldsymbol{z}|\boldsymbol{\mu}, \sigma^2)\, d\mathcal{M}(\boldsymbol{z})$

$$\boldsymbol{z} = \exp_{\boldsymbol{\mu}}^c\left(G(\boldsymbol{\mu})^{-\frac{1}{2}}\, \boldsymbol{v}\right) = \exp_{\boldsymbol{\mu}}^c\left(\boldsymbol{v}/\lambda_{\boldsymbol{\mu}}^c\right) \qquad (2)$$

Isotropic: $\boldsymbol{v} = r\, \boldsymbol{\alpha}$, with direction $\boldsymbol{\alpha} \sim \mathcal{U}(\mathbb{S}^{d-1})$, and hyperbolic radius $r = d_p^c(\boldsymbol{\mu}, \boldsymbol{x})$ density

$$\rho^{\mathsf{R}}(r) \propto 1_{\mathbb{R}_+}(r) e^{-\frac{r^2}{2\sigma^2}} \left(\frac{\sinh(\sqrt{c}r)}{\sqrt{c}}\right)^{d-1} \xrightarrow[c \to 0]{} \rho^{\mathsf{W}}(r) \propto 1_{\mathbb{R}_+}(r)\, e^{-\frac{r^2}{2\sigma^2}}\, r^{d-1} \qquad (3)$$



$\boldsymbol{\sigma} = (1.0, 1.0)$ $\boldsymbol{\sigma} = (2.0, 0.5)$ $\boldsymbol{\sigma} = (0.5, 2.0)$

Figure 2: Wrapped normal probability measures for Fréchet means $\boldsymbol{\mu}$, concentrations $\Sigma = \mathrm{diag}(\boldsymbol{\sigma})$ and $c = 1$.

Decoder & encoder architectures

Decoder Compute geodesic distance to hyperplanes (i.e. *gyroplanes*)

$$f_{\boldsymbol{a},\boldsymbol{p}}(\boldsymbol{z}) = \langle \boldsymbol{a}, \boldsymbol{z} - \boldsymbol{p} \rangle = \mathrm{sign}\left(\langle \boldsymbol{a}, \boldsymbol{z} - \boldsymbol{p} \rangle\right) \|\boldsymbol{a}\|\, d_E(\boldsymbol{z}, H_{\boldsymbol{a},\boldsymbol{p}}^c), \text{ with, } H_{\boldsymbol{a},\boldsymbol{p}} = \boldsymbol{p} + \{\boldsymbol{a}\}^{\perp}.$$

$$f_{\boldsymbol{a},\boldsymbol{p}}^c(\boldsymbol{z}) = \mathrm{sign}\left(\left\langle \boldsymbol{a}, \log_{\boldsymbol{p}}^c(\boldsymbol{z}) \right\rangle_{\boldsymbol{p}}\right) \|\boldsymbol{a}\|_{\boldsymbol{p}}\, d_p^c(\boldsymbol{z}, H_{\boldsymbol{a},\boldsymbol{p}}^c), \text{ with, } H_{\boldsymbol{a},\boldsymbol{p}}^c = \exp_{\boldsymbol{p}}^c(\{\boldsymbol{a}\}^{\perp}) \text{ [8].}$$

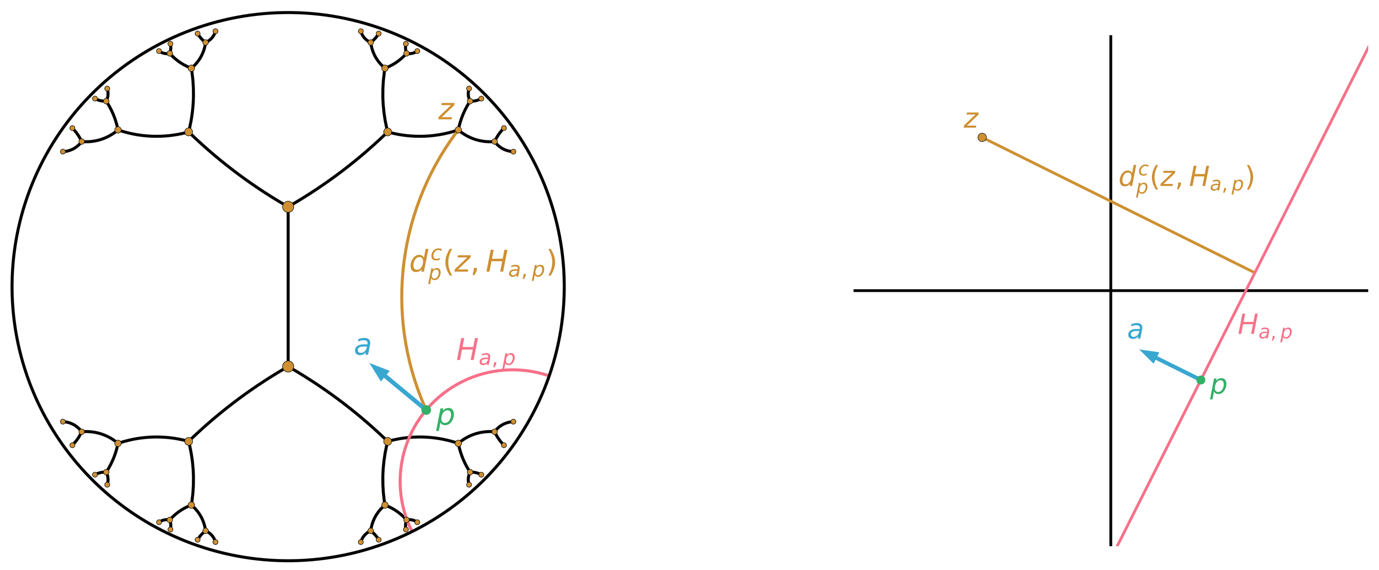


Figure 3: Orthogonal projection on a hyperplane in $\mathbb{B}_c^2$ (a) and $\mathbb{R}^2$ (b).

Encoder $\boldsymbol{\mu} = \exp_{\boldsymbol{\mu}}(\mathrm{enc}_\phi^{\boldsymbol{\mu}}(\boldsymbol{x})) \in \mathbb{B}_c^d$ and $\sigma = \mathrm{softplus}(\mathrm{enc}_\phi^\sigma(\boldsymbol{x})) \in \mathbb{R}_*^+$.

Training

Model $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = p(\boldsymbol{x}|\mathrm{dec}_\theta(\boldsymbol{z}))$, $p(\boldsymbol{z}) = \mathcal{N}_{\mathbb{B}_c^d}(\boldsymbol{z}|\boldsymbol{0}, \sigma_0^2)$ and $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}_{\mathbb{B}_c^d}(\boldsymbol{z}|\mathrm{enc}_\phi^{\boldsymbol{\mu}}(\boldsymbol{x}), \mathrm{enc}_\phi^\sigma(\boldsymbol{x})^2)$.

ELBO $\log p(\boldsymbol{x}) \geq \mathcal{L}_{\mathcal{M}}(\boldsymbol{x}; \theta, \phi) \triangleq \int_{\mathcal{M}} \ln\left(\frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right) q_\phi(\boldsymbol{z}|\boldsymbol{x})\, d\mathcal{M}(\boldsymbol{z})$ via Monte Carlo (4)

Gradients $\nabla_\mu \boldsymbol{z}$ via reparametrisation. $\nabla_\sigma \boldsymbol{z}$ via reparametrisation for the *wrapped* normal and via implicit reparametrisation [9] of ρ^{R} via its cdf $F^{\mathsf{R}}(r; \sigma)$ for the *Riemannian* normal.

Branching diffusion process

Nodes $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) \in \mathbb{R}^n$ are hierarchically sampled as follow

$$\boldsymbol{x}_i \sim \mathcal{N}\left(\cdot\,|\boldsymbol{x}_{\pi(i)}, \gamma^2\right) \quad \forall i \in 1, \dots, N, \;\pi(i) : \text{parent of ith node}$$

Table 1: Negative test marginal likelihood estimates $\mathcal{L}_{\text{IWAE}}$ (with 5000 samples).

	σ_0	Models $\mathcal{N}$-VAE	$\mathcal{P}^{0.1}$-VAE	$\mathcal{P}^{0.3}$-VAE	$\mathcal{P}^{0.8}$-VAE	$\mathcal{P}^{1.0}$-VAE	$\mathcal{P}^{1.2}$-VAE
$\mathcal{L}_{\text{IWAE}}$	1	$57.1_{\pm 0.2}$	$57.1_{\pm 0.2}$	$57.2_{\pm 0.2}$	$56.9_{\pm 0.2}$	$56.7_{\pm 0.2}$	$56.6_{\pm 0.2}$
$\mathcal{L}_{\text{IWAE}}$	1.7	$57.0_{\pm 0.2}$	$56.8_{\pm 0.2}$	$56.6_{\pm 0.2}$	$55.9_{\pm 0.2}$	$55.7_{\pm 0.2}$	$\mathbf{55.6}_{\pm 0.2}$

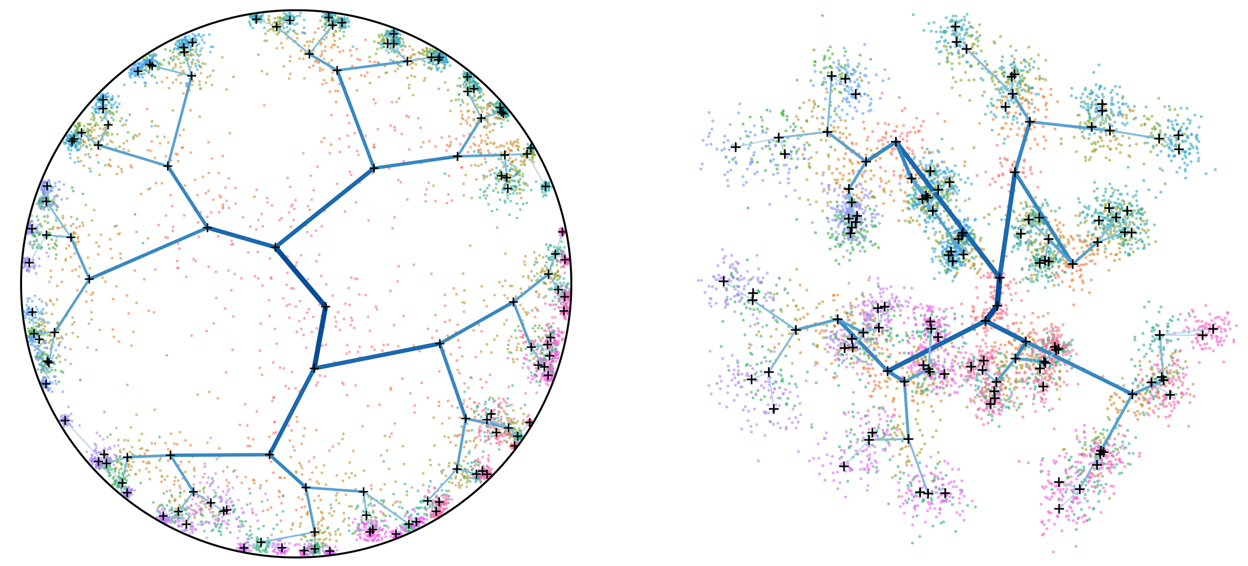


Figure 4: Latent representations learned by $\mathcal{P}^1$-VAE (a), $\mathcal{N}$-VAE (b). Embeddings are represented by black crosses, and colour dots are posterior samples. Blue lines represent true hierarchy.

MNIST digits

One can view the natural clustering in MNIST images as a hierarchy with each of the 10 classes being internal nodes of the hierarchy.

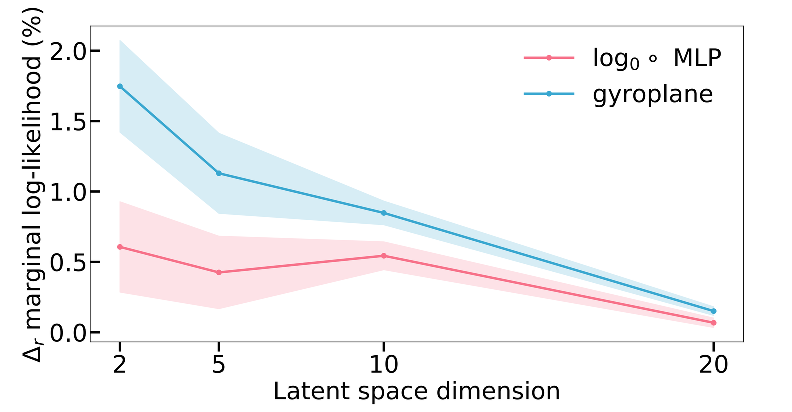


Figure 5: Decoder ablation study with *wrapped* normal $\mathcal{P}^1$-VAE. Baseline decoder is a MLP. We additionally compare the *gyroplane* layer introduced in the Decoder Section to a MLP pre-composed by $\log_0$.

Table 2: Per digit accuracy of a classifier trained on the 2-d latent embeddings. Results are averaged over 10 sets of embeddings and 5 classifier trainings.

Digits	0	1	2	3	4	5	6	7	8	9	Avg
$\mathcal{N}$-VAE	89	97	81	75	59	43	89	**78**	68	**57**	73.6
$\mathcal{P}^{1.4}$-VAE	**94**	97	**82**	**79**	**69**	**47**	**90**	77	68	53	**75.6**

Graph embeddings

We evaluate the performance of a variational graph auto-encoder (VGAE) [10] for link prediction in networks.

Table 3: Results on network link prediction. 95% confidence intervals are computed over 40 runs.

	Phylogenetic		CS PhDs		Diseases	
	AUC	**AP**	**AUC**	**AP**	**AUC**	**AP**
$\mathcal{N}$-VAE	$54.2_{\pm 2.2}$	$54.0_{\pm 2.1}$	$56.5_{\pm 1.1}$	$56.4_{\pm 1.1}$	$89.8_{\pm 0.7}$	$91.8_{\pm 0.7}$
$\mathcal{P}$-VAE	$\mathbf{59.0}_{\pm \mathbf{1.9}}$	$55.5_{\pm 1.6}$	$\mathbf{59.8}_{\pm 1.2}$	$56.7_{\pm 1.2}$	$\mathbf{92.3}_{\pm 0.7}$	$\mathbf{93.6}_{\pm 0.5}$