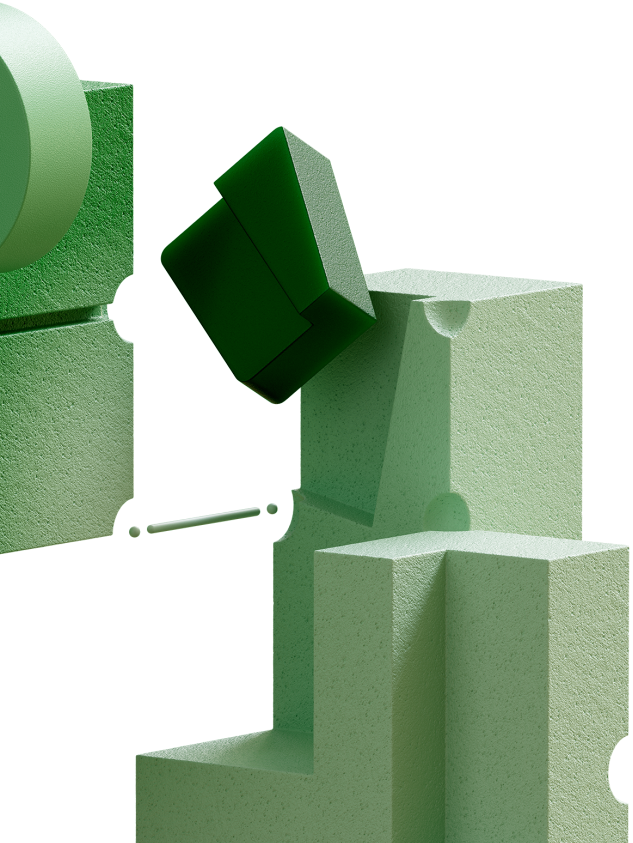
Google DeepMind Inria AISTATS



Beyond Binning: Soft Task Reformulation for Deep Regression

Lawrence Stewart

Francis Bach

Quentin Berthet

Regression and Classification

Features x and targets y

Prediction with $z = f_\eta(x)$

Minimizing MSE (or other loss)

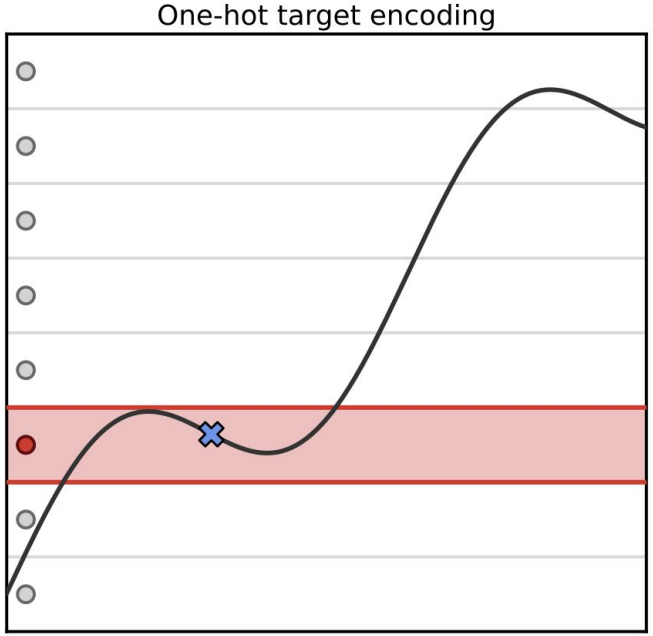
$$\min_\eta \ \mathbf{E}_{(x,y)}\left[L(y - f_\eta(x))\right]$$

Suboptimal, suffers from implicit biases

<u>Hard binning classification</u> $\psi_c(y) = e_i$

$$\min_\theta \mathbf{E}[\mathrm{KL}(\psi_w(y) \| \pi_\theta(x))]$$!?

(Stewart et al. 2023, Farebrother et al. 2024, Grinsztajn et al. 2022, Chizat and Bach 2018, Pintea et al. 2023, Imani and White 2018)



[NEW] - Soft binning

Soft binning $\psi_c(y) = \mathrm{softmax}(\|y - c_i\|^2)$

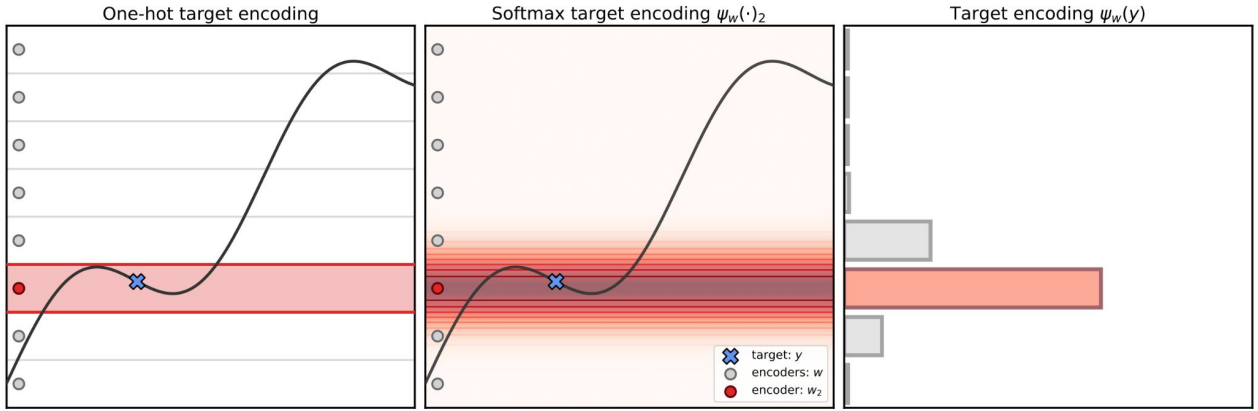
Probabilistic target encoding, differentiable - prediction $\hat{y}_{\text{pred}} = c^\top \pi_\theta(x)$

Classification objective $\min_\theta \mathbf{E}[\mathrm{KL}(\psi_c(y) \| \pi_\theta(x))]$

Pros:

Reduce information loss, quantization.

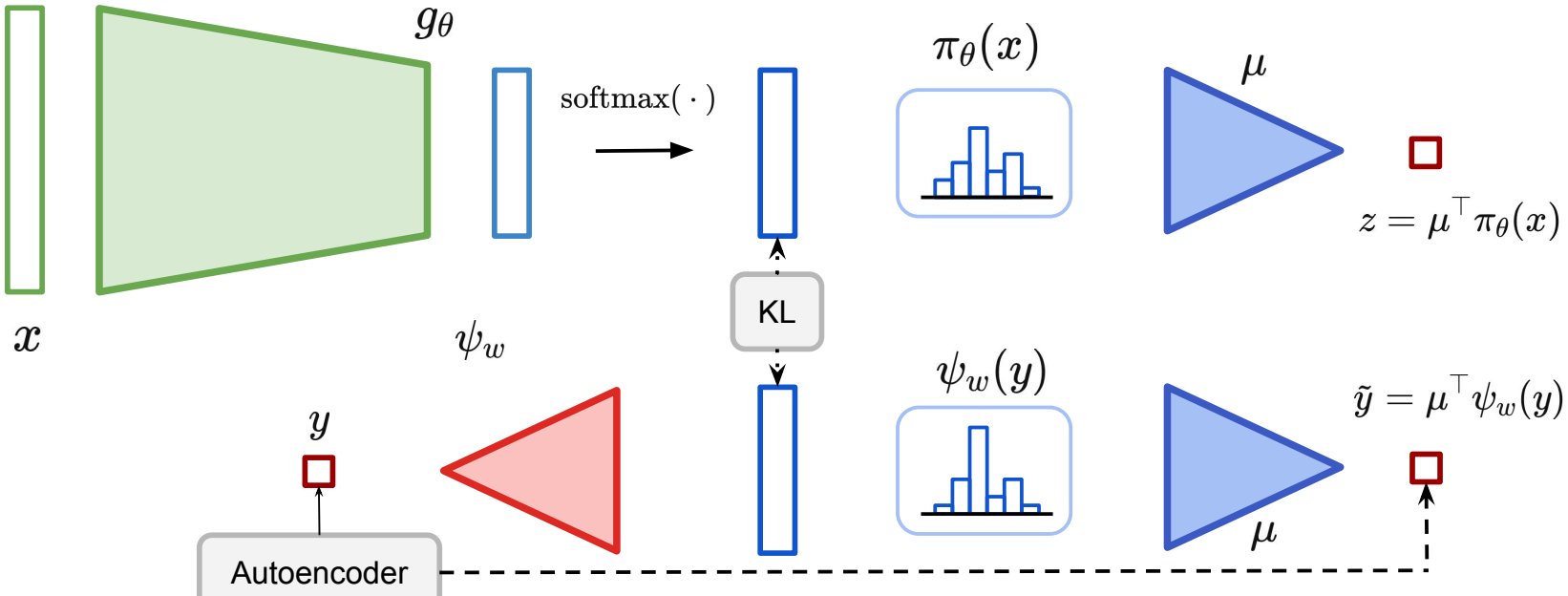
General encodings?



[NEW] - Trained autoencoder

General encoder: $\psi_w(y) = \mathrm{softmax}(w_1^\top y + y_2)$ - Decoder: $\mu^\top \pi_\theta(x)$

Autoencoding loss: $\min_{w,\mu} \mathbf{E}_y\left[L(y - \mu^\top \psi_w(y))\right]$



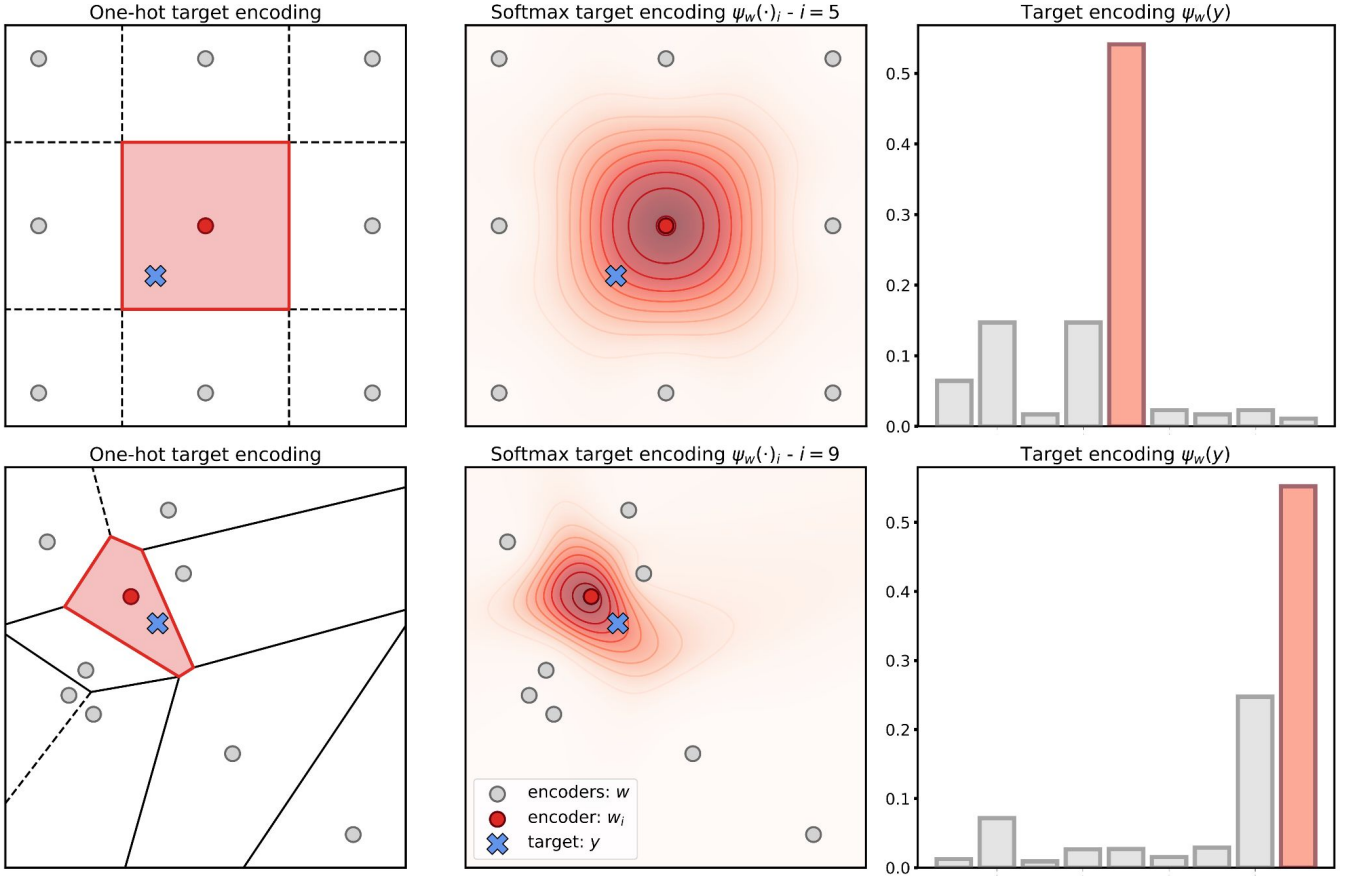
[NEW] - Trained autoencoder

Generalizable encodings

Adapted to data distribution

Illustrated here for 2D targets

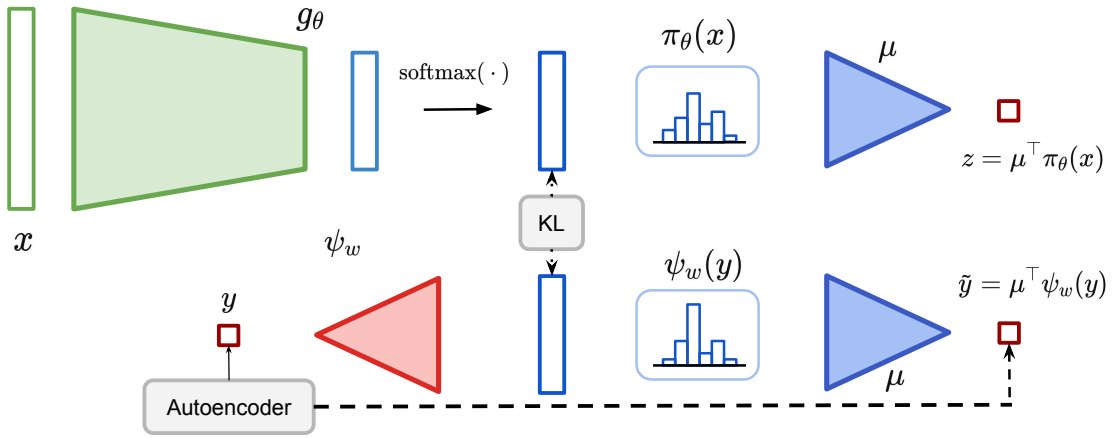
Training it to be classification-aware



[NEW] - Bridging objectives - end-to-end loss

Mix all objectives

- Autoencoding loss
- Classification loss
- Prediction loss



Bridges these problems: regression in probability space
Encoding with downstream objective, single joint objective to minimize

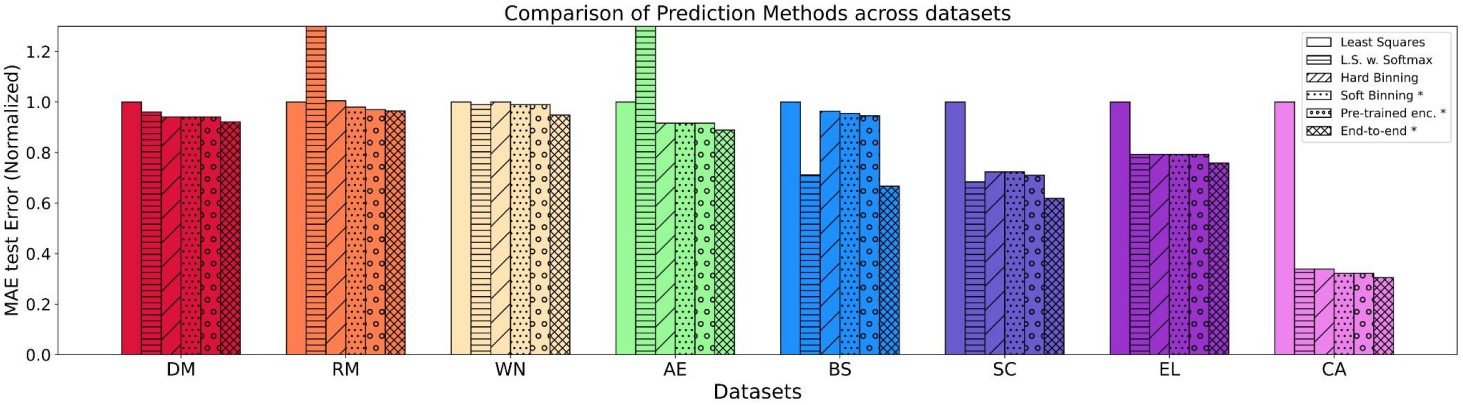
$$\min_{w,\mu,\theta} \lambda_{\text{auto}} \mathbf{E}_y[L(y - \mu^\top \psi_w(y))] + \lambda_{\text{KL}} \mathbf{E}_{(x,y)}[KL(\psi_w(y)||\pi_\theta(x))]$$
$$+ \lambda_{\text{pred}} E_{(x,y)}[L(y - \mu^\top \pi_\theta(x))]$$

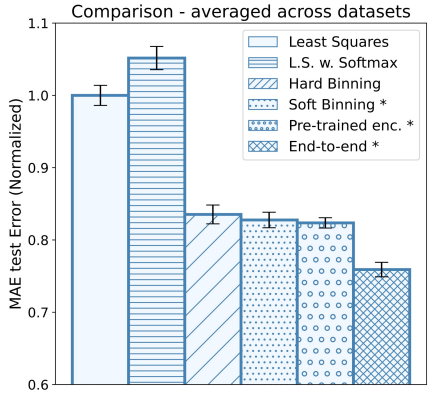
Experimental results

Comparison of these approaches on standard open regression tasks

- Evaluated on 8 real-world datasets, Test-time MAE, rMSE, and R^2.
- Global Hierarchy & Ablation study: soft binning and trained auto-encoder help

<u>End-to-End Advantage</u>: Main proposed objective has best performance across all datasets bridging classification and regression

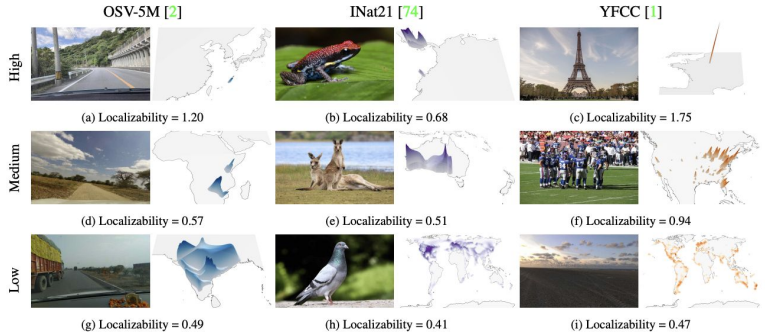




Experimental results

Large-scale experiment

- Open Street View 5M (OSV5M) featuring 5.1 million georeferenced images.
- Model: ViT-L Transformer with a 12-layer MLP head.
- 30% reduction in MAE and 7.5% increase in R^2 wrt MSE baseline.



Metric	Least-squares	End-to-end	Δ
MAE (↓)	0.073 ± 0.002	0.051 ± 0.002	-30.04%
rMSE (↓)	0.114 ± 0.003	0.099 ± 0.000	-13.04%
$\mathbf{R}^2$ (↑)	0.749 ± 0.006	0.806 ± 0.001	+7.52%

Table 2: Open Street View 5M: test set statistics

(Dufour et al. 2024)