Interpreting Uncertainty in Bayesian Neural Networks

UNIVERSITY OF CAMBRIDGE

Javier Antorán, ja666@cam.ac.uk

Bayesian Neural Networks (BNN)

- Parameter point estimates are substituted by probability distributions.
- Automatically balance goodness of fit and model simplicity.
- Uncertainty in the weights is translated into uncertainty in predictions.

Can uncertainty in BNNs help us better understand our data?

SG-HMC Approximate Inference

The true posterior over network weights, $p(\mathbf{w}|\mathcal{D})$, is intractable. Approximate inference methods tend to present a trade-off between fitting the data well and providing reliable uncertainty estimates. Drawing samples with SG-HMC allows for both, [2].

Hamiltonian Monte Carlo introduces an auxiliary momentum variable $\mathbf{r}$ and samples from the joint distribution: $p(\mathbf{w}, \mathbf{r}|\mathcal{D}) \propto \exp\left(\log p(\mathcal{D}, \mathbf{w}) - \frac{1}{2}\mathbf{r}\mathbf{M}\mathbf{r}^\intercal\right)$. Stochastic approximations to gradients are computed using minibatches $\widetilde{\mathcal{D}}$. A friction term $\mathbf{C}$ is used to compensate for the variance introduced by this stochasticity.

$$\begin{cases} \Delta\mathbf{w} = \epsilon\mathbf{M}^{-1}\mathbf{r} \\ \Delta\mathbf{r} = \epsilon\nabla_{\mathbf{w}}\log p(\mathbf{w}, \widetilde{\mathcal{D}}) - \epsilon\mathbf{C}\mathbf{M}^{-1}\mathbf{r} + \gamma; \quad \gamma \sim \mathcal{N}(0,\, 2(\mathbf{C} - \mathbf{B})\cdot\epsilon) \end{cases}$$

The mass matrix $\mathbf{M}$ and gradient noise $\mathbf{B}$ are estimated during burn-in. Expectations of functions parametrised by $\mathbf{w}$ can be approximated as:

$$\mathbb{E}_{p(\mathbf{w})}[f_{\mathbf{w}}] \approx \frac{1}{N}\sum_{i=1}^{N} f_{\mathbf{w}^{(i)}}$$

Decomposing Uncertainty

Irreducible or **Aleatoric uncertainty** can be quantified as $\mathcal{H}_a = \mathbb{E}_{p(\mathbf{w})}[\mathcal{H}(\mathbf{y}\,|\,\mathbf{x}, \mathbf{w})]$. It indicates that there are unobserved factors which influence our targets or that our measurements are noisy. Model uncertainty or **Epistemic uncertainty** can be measured as $\mathcal{H}_e = \mathcal{H}(\mathbf{y}\,|\,\mathbf{x}) - \mathcal{H}_a$. It can be reduced by observing more data, [1].

For a heteroscedastic regression scenario with a Gaussian distribution over outputs, predictive variance can be decomposed as:

$$Var(y|\mathbf{x}) \approx \underbrace{\frac{1}{N}\sum_{i=1}^{N}\mu_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\mu_i^2\right)}_{\text{epistemic}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2}_{\text{aleatoric}}$$

References

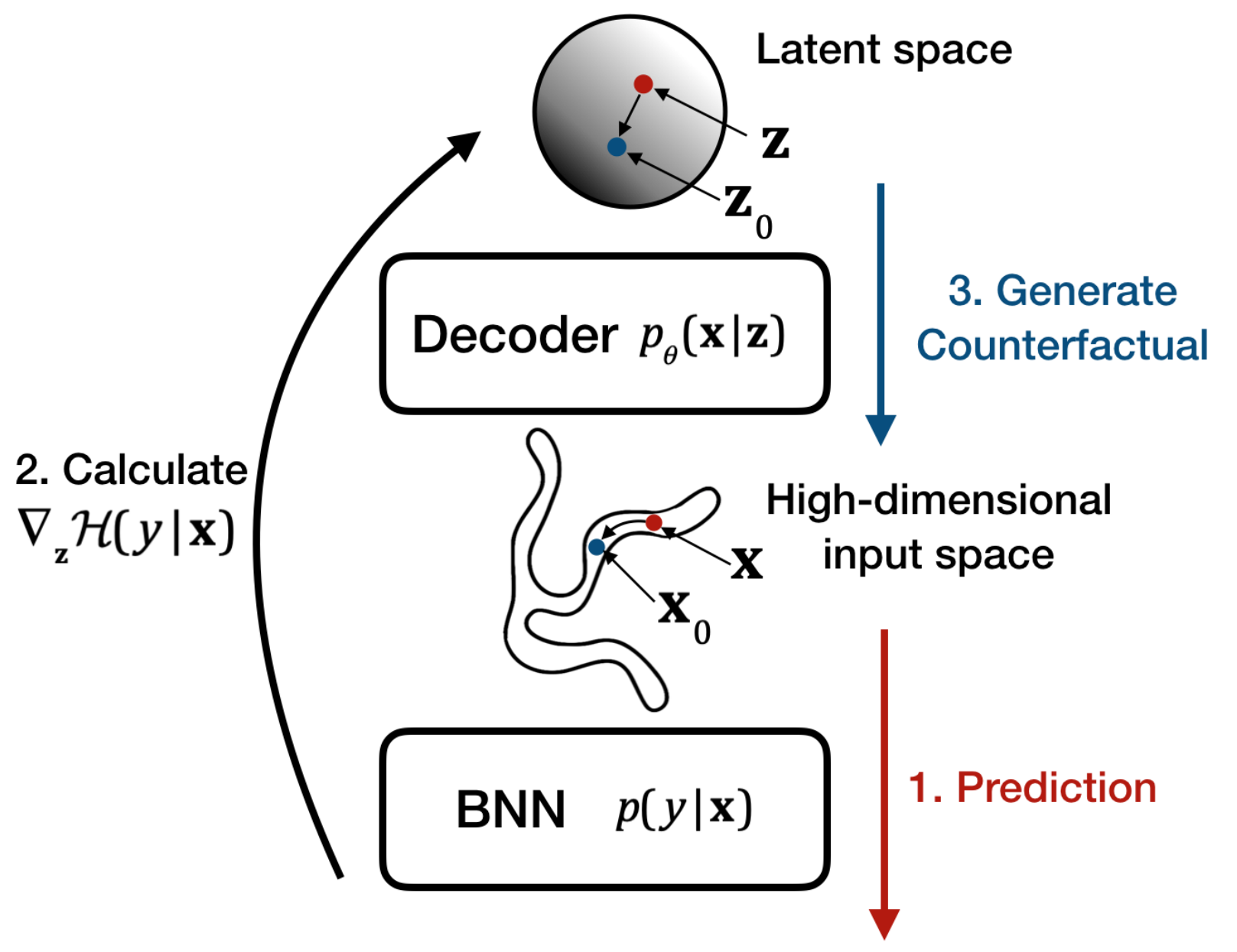
[1] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning.

[2] J. T. Springenberg, A. Klein, S. Falkner, and F. Hutter. Bayesian optimization with robust bayesian neural networks.

Explaining Uncertainty with Counterfactuals

For input point $\mathbf{x}$, we can answer the question: **"Why is this datapoint aleatoric or epistemic?"** with a counterfactual example. We generate the counterfactual $\mathbf{x}_0$ using gradient descent with respect to the entropy. A Variational Autoencoder (VAE) is used to constrain the space of explanations: $\mathbf{x}_0 = \mathbb{E}[\mathbf{x}|\mathbf{z}_0]$:

$$\Delta\mathbf{z}_e = \frac{\partial}{\partial\mathbf{z}}\mathcal{H}_e; \quad \Delta\mathbf{z}_a = \frac{\partial}{\partial\mathbf{z}}\mathcal{H}_a;$$



Latent space
$\mathbf{z}$
$\mathbf{z}_0$
Decoder $p_\theta(\mathbf{x}|\mathbf{z})$
3. Generate Counterfactual
2. Calculate $\nabla_{\mathbf{z}}\mathcal{H}(y|\mathbf{x})$
High-dimensional input space
$\mathbf{x}$
$\mathbf{x}_0$
BNN $p(y|\mathbf{x})$
1. Prediction

Figure: Gradients from the BNN are propagated back to the latent space. Steps are taken towards less entropic latent regions.

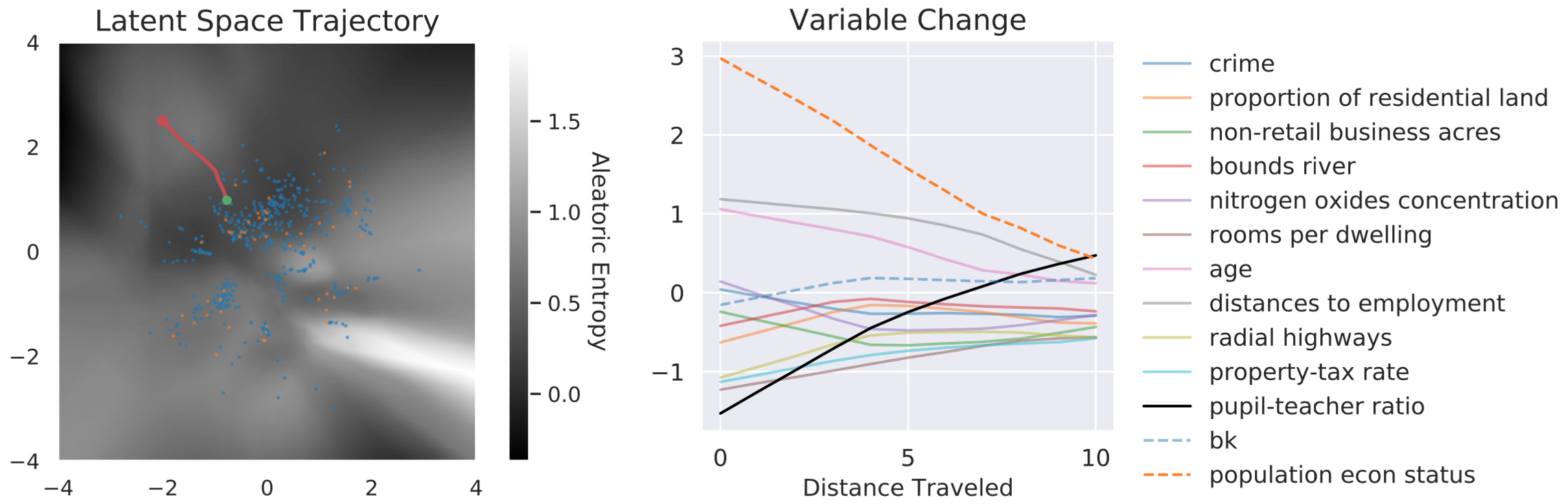


Figure: Latent trajectory (left) and normalised input space explanation (right) for the aleatoric entropy of a single sample from the Boston housing dataset. $\Delta\mathcal{H}_a = -1.2$.

In the above example, two factors are principally responsible for the large aleatoric entropy: the high the economic status of the population and the low pupil-teacher ratio. Seeing this, we could draw the tentative conclusion that birth rates should also be used as an input variable.

Comparison with Uncertainty Sensitivity Analysis

Sensitivity analysis computes gradients in input space: $S_a(\mathbf{x}) = \left(\frac{\partial}{\partial\mathbf{x}}\mathcal{H}_a\right)^2$.

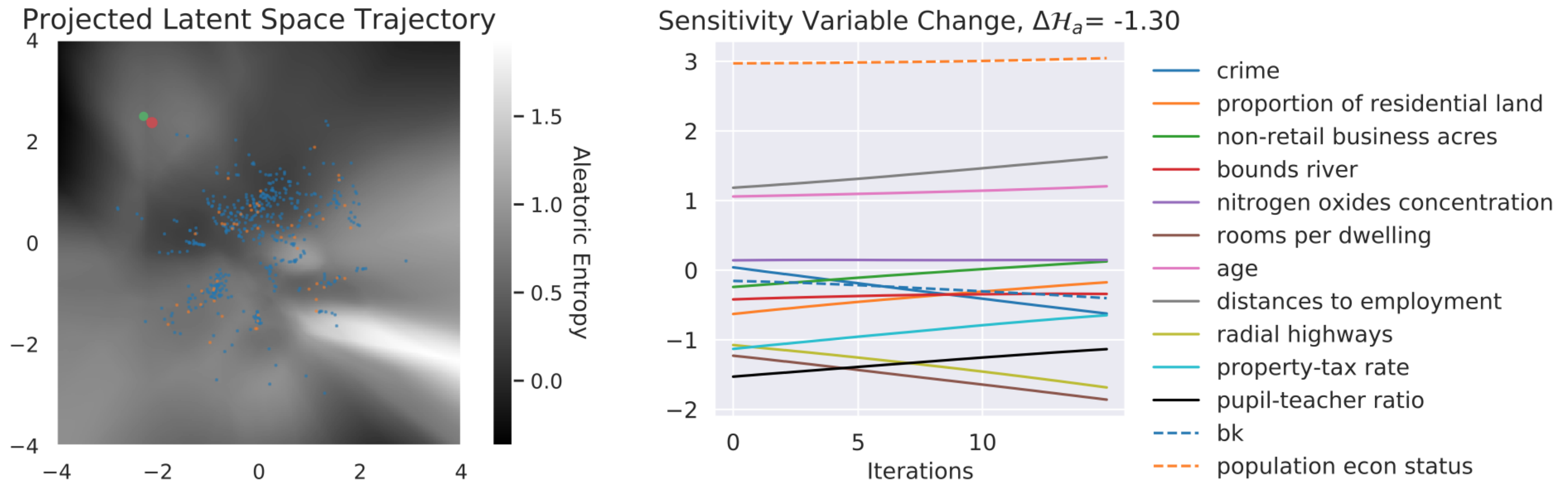


Figure: Following the gradient in input space results in infinitesimal variations in latent space (left). Small changes in input space result in a large entropy decrease (right).

Small changes in input space can greatly reduce $\mathcal{H}_a$. This is analogous to creating an adversarial example. The proposed method's use of a generative model ensures that our explanations are in-distribution.

Going from Local to Global Explanations

K-means clustering is applied to $\{\Delta\mathbf{x}\} = \{\mathbf{x}_0 - \mathbf{x}\}$. Each cluster's mean $\Delta\mathbf{x}$ acts as a "prototypical point". The whole dataset is explained through the uncertainty of a finite set of **"prototypical points."**

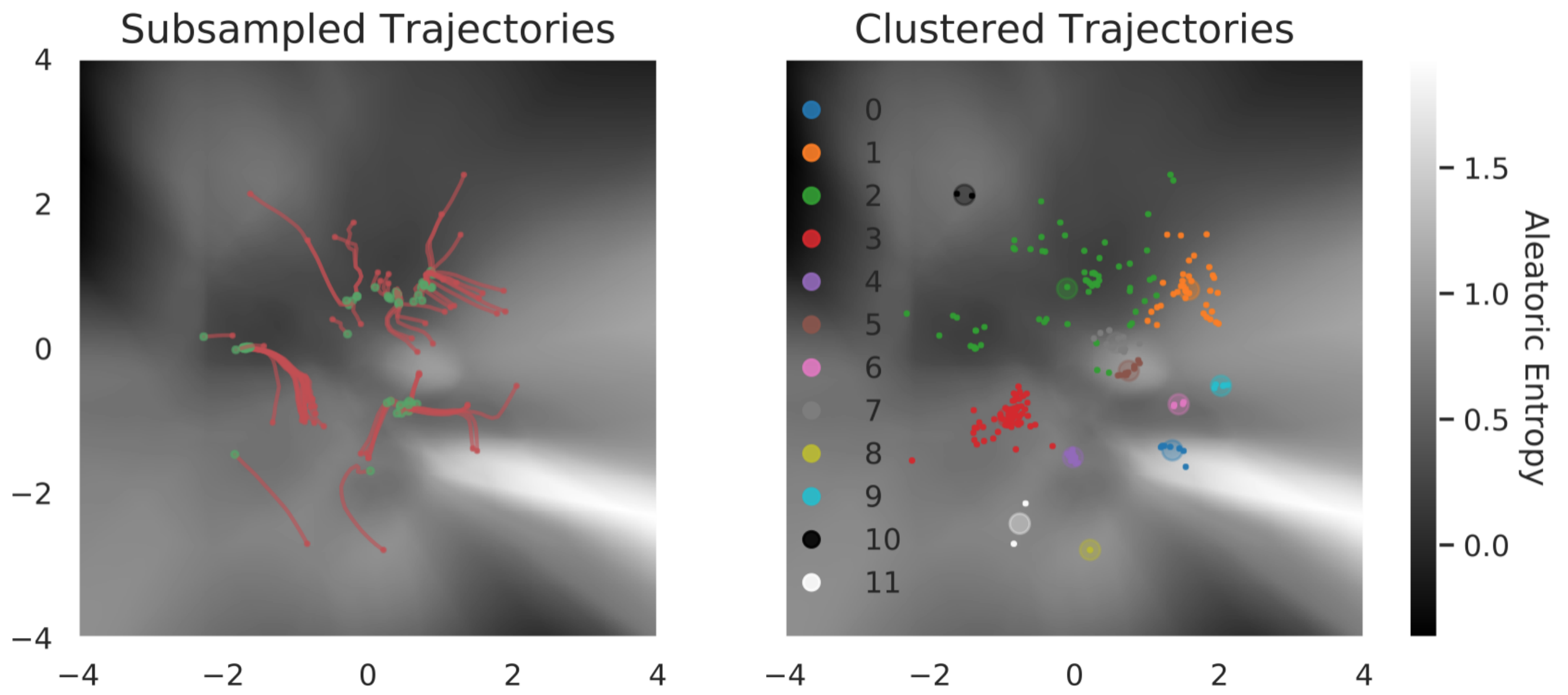


Figure: Subsample of all points' latent trajectories (left) and latent projections of clustered points (right). Clustering distances are based on $\Delta\mathbf{x}$, in input space.

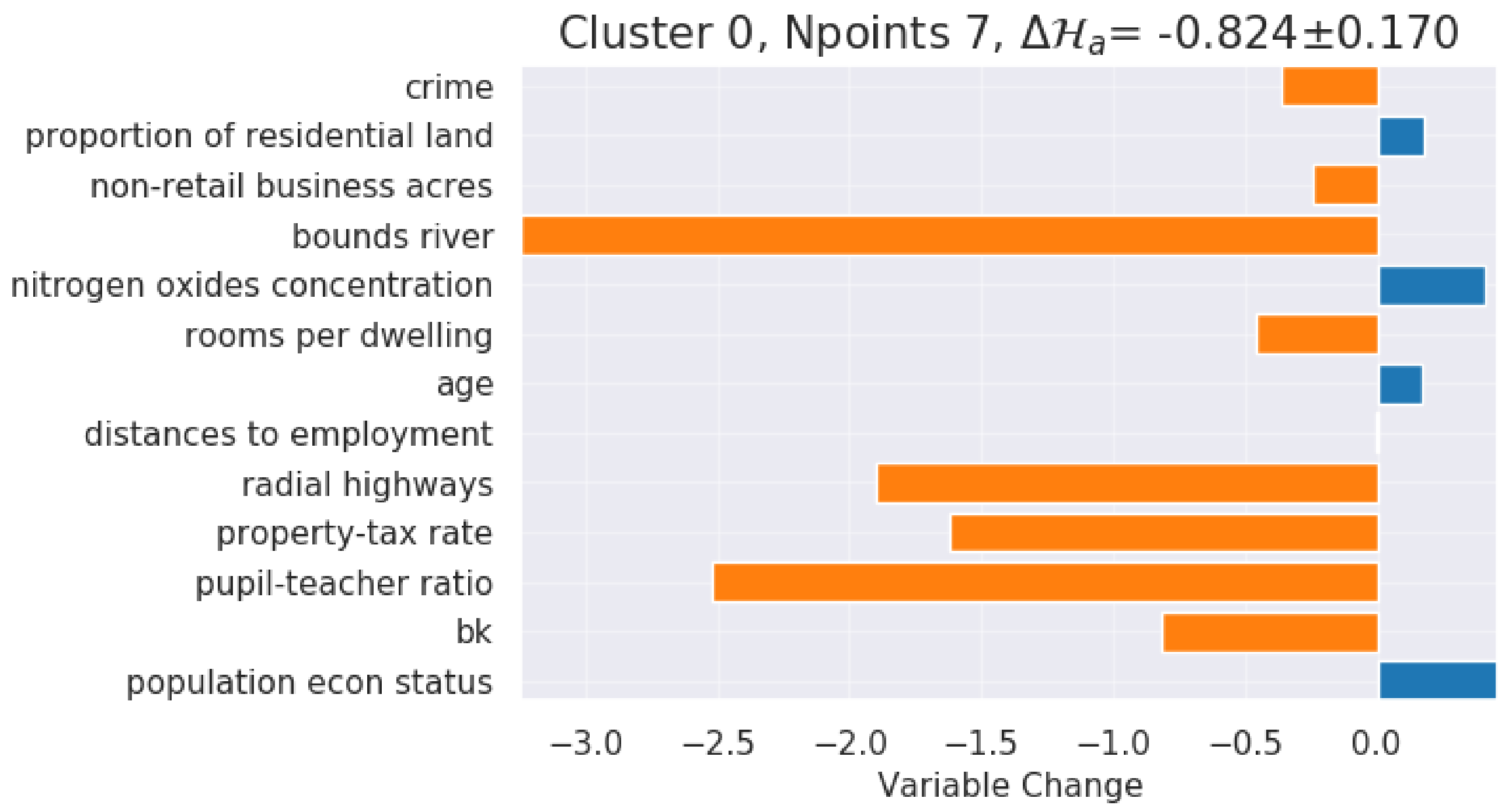


Figure: Mean input variable changes for points in cluster 0. Uncertainty stems mainly from: bounding rivers, radial highways, pupil teacher ratio and property tax rate.

Explaining Uncertainty in Images

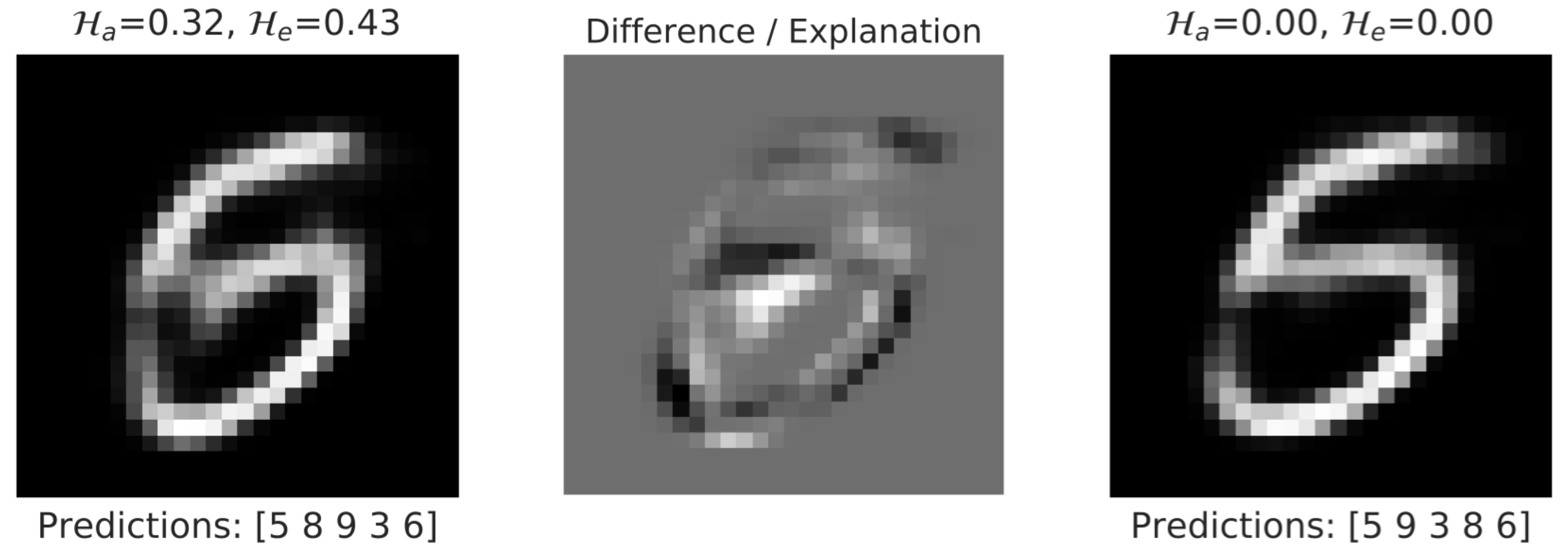


Figure: A digit with high uncertainty (left). The explanation given by our method (center). The same digit with reduced uncertainty (right). Note how the second most probable class, 8, becomes less likely after the explanation is subtracted. The explanation's positive values resemble an 8 while its negative values resemble a 5.