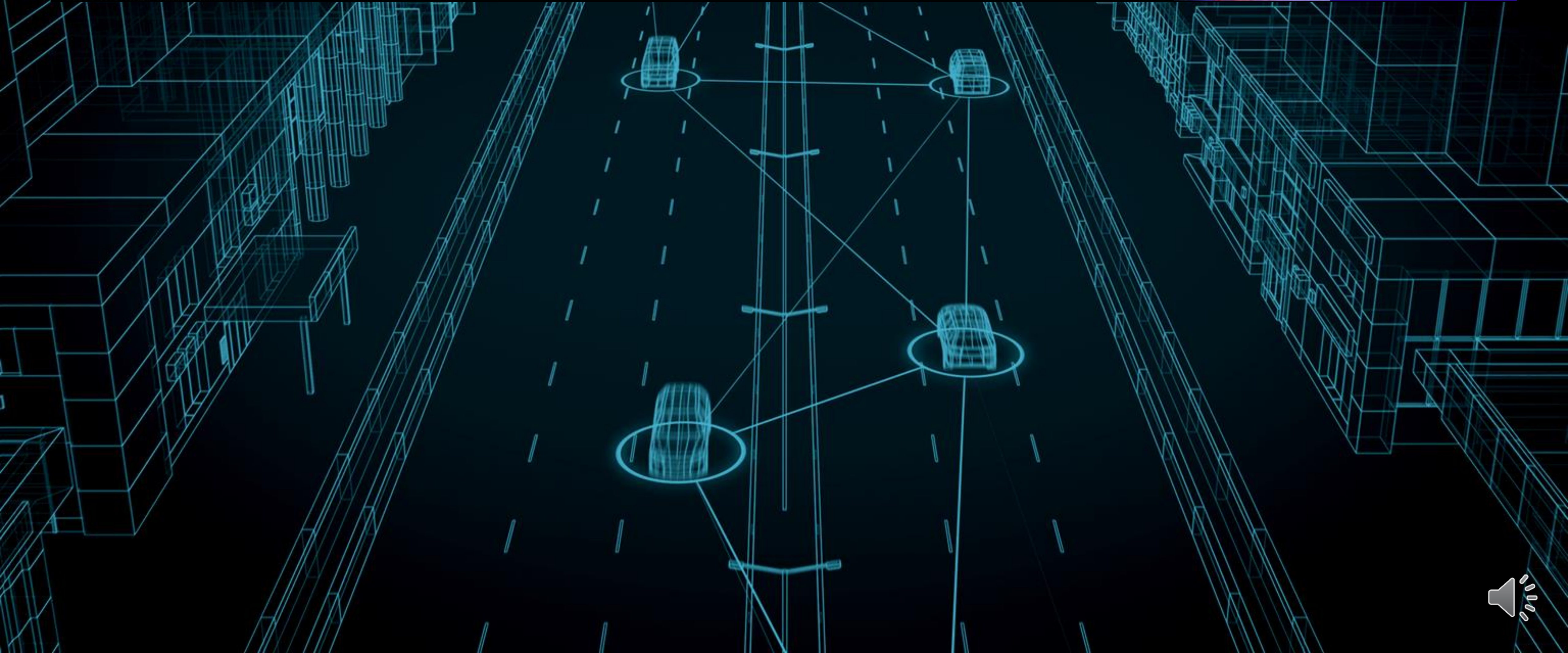
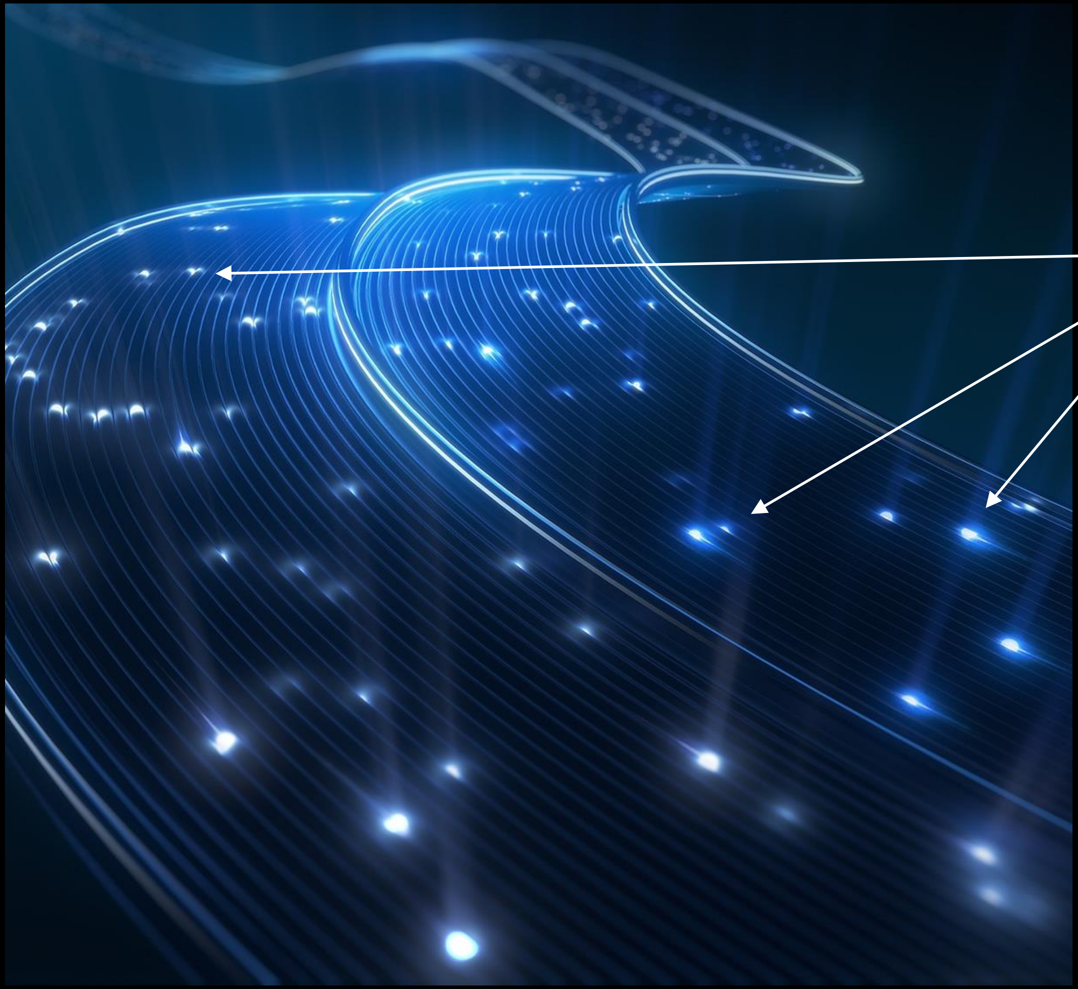
# Bayesian Uncertainty and Expected Gradient Length - Regression: Two Sides of The Same Coin?

Megh Shukla
Mercedes-Benz Research and Development India



# Information Superhighway



Labelling all this data is *very, very* expensive

Active Learning: Can we select images to label such that we…

***Maximize model performance per set of images annotated***

# Expected Gradient Length (EGL)

Gradient as a measure of informative content for an image

Classification [1] : $x^* = \arg\max \sum_i p(y_i|x,\theta) \, ||\nabla_\theta l(\mathcal{L} \cup \{x, y_i\}; \theta)||$

Regression [2] : $x^* = \arg\max \frac{1}{K} \sum_{k=1}^{K} ||(f_z(x) - f_k(x))x||$

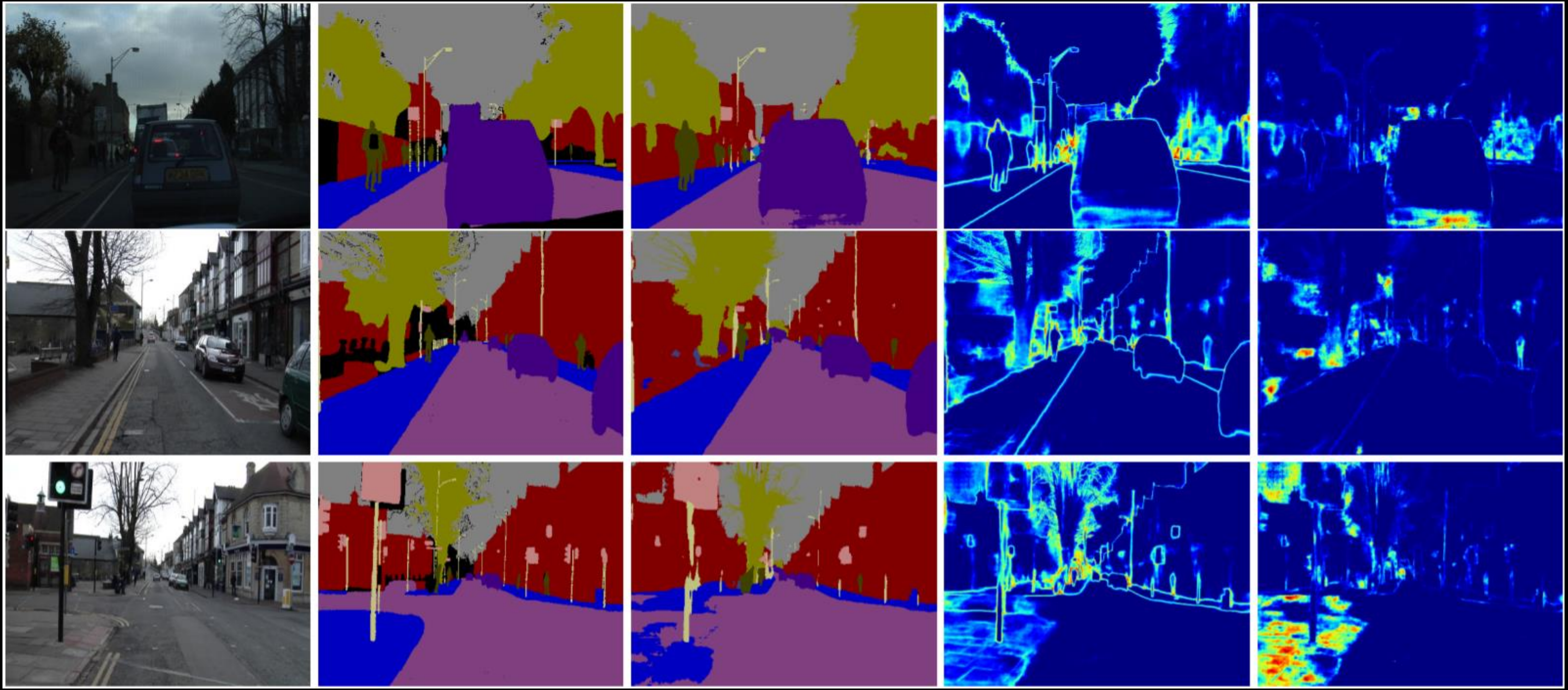
Intuitively defined, lacks theory! ☹

[1] Settles and Craven, *"An Analysis of Active Learning Strategies for Sequence Labeling Tasks"*, EMNLP 2008

[2] Cai et al., *"Maximizing expected model change for active learning in regression"*, ICDM 2013

# Bayesian Uncertainty in Computer Vision [3]

$$Var(y) \approx \underbrace{\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t\right)^2}_{\text{Epistemic}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_t^2}_{\text{Aleatoric}}$$



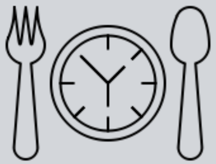
Aleatoric Epistemic

[3] Kendall and Gal, “*What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?*”, NeurIPS 2017

# Are Bayesian Uncertainty and Expected Gradient Length equivalent?

Literature places uncertainty and expected gradient length as different active learning paradigms. Is there a connection between them?

SPOILER: Yes ☺



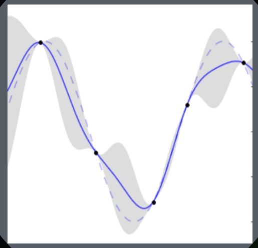
The No Free Lunch Theorem

Fisher Information

$$\int_{\mathbb{R}} \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,dx$$

Linear Regression

Non-linear Regression



# The No Free Lunch Theorem [4] … But Why?

$\mathcal{Z} \in \{\mathcal{X} \times \mathcal{Y}\}$ — Sample space

$p(x, y|\theta_0) = p(y|x, \theta_0)p(x)$ — True distribution

$q(x, y|\theta_0) = p(y|x, \theta_0)q(x)$ — Sample distribution


*There is no universal learner, no learner can succeed on all learning tasks*

[4] Shalev-Shwartz and Ben David, *"Understanding Machine Learning: From Theory to Algorithms"*, Cambridge Press

# The No Free Lunch Theorem [4] … But Why?

*There is no universal learner, no learner can succeed on all learning tasks*

$$q(x, y|\theta_0) = p(y|x, \theta_0)q(x)$$

Bayesian Analysis

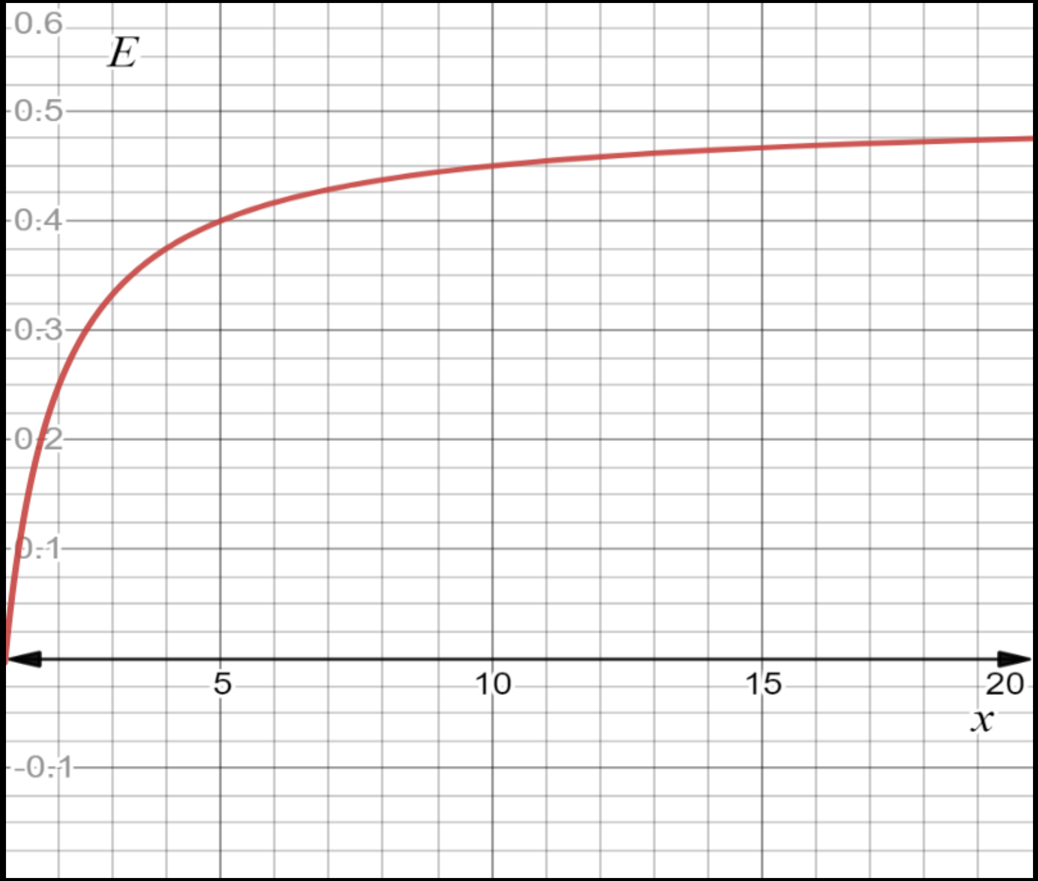
$$q(x, y) = \int_\theta q(y|x, \theta)q(x)\pi(\theta|z_{obs})\mathrm{d}\theta$$

But do we need to do this … ?

*YES.*

Namely, let $A$ be a learning algorithm for the task of binary classification. Let $m$ be any number smaller than $|\mathcal{X}|/k$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ such that:

- There exists a function $f : \mathcal{X} \to \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$.
- $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \frac{1}{2} - \frac{1}{2k}$.




Notice the convergence of expected risk to chance as the ratio of unlabelled points increases

[4] Shalev-Shwartz and Ben David, *"Understanding Machine Learning: From Theory to Algorithms"*, Cambridge Press

# Fisher Information

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} \mathcal{N}(0, I_q^{-1}(\theta_0)))$$

*Distribution*

$$q(x, y) = \int_\theta q(y|x, \theta) q(x) \pi(\theta | z_{obs}) \mathrm{d}\theta$$

*Substitution*

$$q^* = \arg\max_q \int_x q(x) \int_y \int_\theta q(y|x,\theta)\pi(\theta|z_{obs}) \|\nabla_{\theta_0} l(x, y, \theta_0)\|^2 \, \mathrm{d}\theta \, \mathrm{d}y \, \mathrm{d}x$$

*Ensemble*

$$x^* = \arg\max \frac{1}{K} \sum_{k=1}^{K} \int_y q(y|x, \theta_k) \|\nabla_{\theta_z} l(x, y, \theta_z)\|^2$$

# Expected Gradient Length - Regression

$$x^* = \arg\max \frac{1}{K} \sum_{k=1}^{K} \int_y q(y|x, \theta_k) \|\nabla_{\theta_z} l(x, y, \theta_z)\|^2$$

$$q(y|x, \theta)$$

$$\|\nabla_{\theta_z} l(x, y, \theta_z)\|^2$$

$$\mathcal{N}(y, \mu_k, \sigma_k)$$

$$(y - f_z(x))\nabla_{\theta_z} f_z(x)$$

CLOSED FORM SOLUTION

# Expected Gradient Length - Regression

## CLOSED FORM SOLUTION

$$||\nabla_{\theta_z} f_z(x)||^2 \; [(y - \mu_k)) + (\mu_k - f_z)]^2$$

$$||\nabla_{\theta_z} f_z(x)||^2 \quad \int \mathcal{N}(y, \mu_k, \sigma_k)(y - \mu_k)^2 \mathrm{d}y + 2 \int \mathcal{N}(y, \mu_k, \sigma_k)(y - \mu_k)(\mu_k - f_z)\mathrm{d}y + \int \mathcal{N}(y, \mu_k, \sigma_k)(\mu_k - f_z)^2 \mathrm{d}y$$

VARIANCE

MEAN – MEAN = 0

$(\mu_k - f_z)^2$

$$x^* = \operatorname{argmax} \frac{||\nabla_{\theta_z} f_z(x)||^2}{K} \sum_{k=1}^{K} \hat{\sigma}_k^2(x) + (\mu_k(x) - f_z(x))^2$$

For reference – Bayesian Uncertainties in Computer Vision

$$Var(y) \approx \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T} \hat{y}_t\right)^2 + \frac{1}{T}\sum_{t=1}^{T} \hat{\sigma}_t^2$$

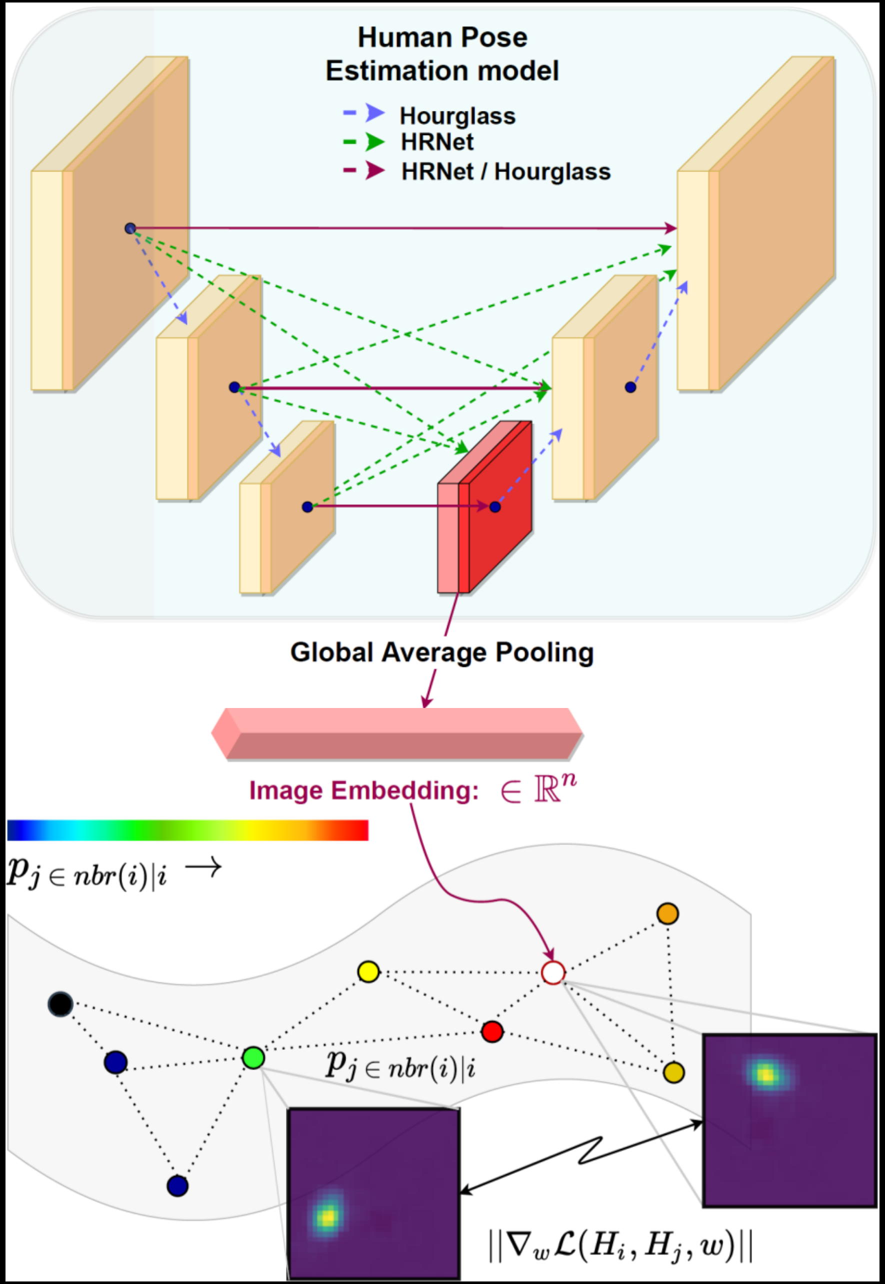
# EGL++

No Ensembles / Dropouts?

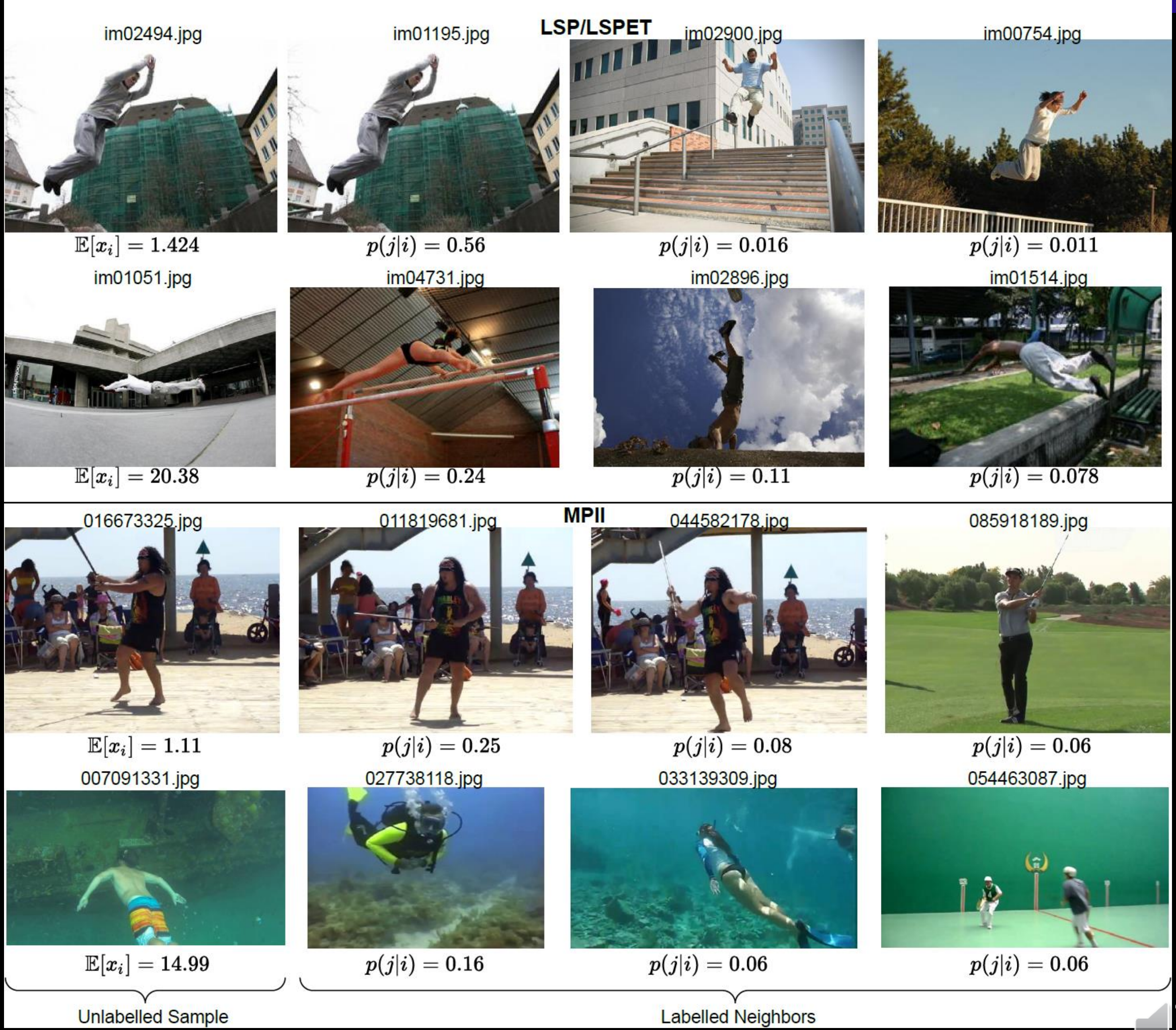
No closed form solution for $q(y|x, \theta)$ ?

$$x^* = \arg\max \frac{1}{K} \sum_{k=1}^{K} \int_y q(y|x, \theta_k) \|\nabla_{\theta_z} l(x, y, \theta_z)\|^2$$

$$x^* = \arg\max \sum_{n=1}^{n=N} q_{tsne}(y_n|x, \theta) \|\nabla_\theta l(x, y_n, \theta)\|^2$$



Human Pose
Estimation model
Hourglass
HRNet
HRNet / Hourglass
Global Average Pooling
Image Embedding: $\in \mathbb{R}^n$
$p_{j \in nbr(i)|i} \rightarrow$
$p_{j \in nbr(i)|i}$
$||\nabla_w \mathcal{L}(H_i, H_j, w)||$

LSP/LSPET
im02494.jpg
im01195.jpg
im02900.jpg
im00754.jpg
$\mathbb{E}[x_i] = 1.424$
$p(j|i) = 0.56$
$p(j|i) = 0.016$
$p(j|i) = 0.011$
im01051.jpg
im04731.jpg
im02896.jpg
im01514.jpg
$\mathbb{E}[x_i] = 20.38$
$p(j|i) = 0.24$
$p(j|i) = 0.11$
$p(j|i) = 0.078$
MPII
016673325.jpg
011819681.jpg
044582178.jpg
085918189.jpg
$\mathbb{E}[x_i] = 1.11$
$p(j|i) = 0.25$
$p(j|i) = 0.08$
$p(j|i) = 0.06$
007091331.jpg
027738118.jpg
033139309.jpg
054463087.jpg
$\mathbb{E}[x_i] = 14.99$
$p(j|i) = 0.16$
$p(j|i) = 0.06$
$p(j|i) = 0.06$
Unlabelled Sample
Labelled Neighbors

# EGL++

MPII Newell Validation Split: Mean+-Sigma (5 runs), one-tailed paired t-test (vs EGL++) at 0.1 significance value

| #images → | 2000 | | | 3000 | | | 4000 | | | 5000 | | | 6000 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | $\mu$ | $\sigma$ | $p-value$ | $\mu$ | $\sigma$ | $p$ | $\mu$ | $\sigma$ | $p$ | $\mu$ | $\sigma$ | $p$ | $\mu$ | $\sigma$ | $p$ |
| Random | 75.95 | 0.55 | **0.003** | 78.33 | 0.65 | **0.012** | 80.31 | 0.91 | **0.006** | 81.35 | 0.41 | **0.001** | 82.23 | 0.74 | **0.007** |
| Core-set [37] | 76.61 | 0.6 | **0.0047** | 79.24 | 0.7 | 0.245 | 81.25 | 0.67 | **0.072** | 82.23 | 1.14 | 0.123 | 82.97 | 1.11 | **0.064** |
| Multi-peak [27] | 76.74 | 0.61 | **0.0054** | 79.56 | 0.46 | 0.462 | 81.19 | 0.31 | **0.063** | 82.61 | 0.5 | **0.093** | 83.11 | 0.71 | **0.062** |
| Learning Loss [51] | 76.28 | 0.76 | **0.0276** | 79.27 | 0.52 | 0.185 | 81.35 | 0.35 | 0.152 | 82.94 | 0.44 | 0.319 | 83.79 | 0.46 | **0.053** |
| EGL++ (ours) | *77.28* | 0.63 | - | *79.58* | 0.33 | - | *81.53* | 0.51 | - | *83.07* | 0.25 | - | *84* | 0.38 | - |

LSP Test Split: Mean+-Sigma (5 runs), one-tailed paired t-test (vs EGL++) at 0.1 significance value

| #images → | 2000 | | | 3000 | | | 4000 | | | 5000 | | | 6000 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | $\mu$ | $\sigma$ | $p-value$ | $\mu$ | $\sigma$ | $p$ | $\mu$ | $\sigma$ | $p$ | $\mu$ | $\sigma$ | $p$ | $\mu$ | $\sigma$ | $p$ |
| Random | 80.34 | 0.31 | 0.285 | 81.81 | 0.24 | 0.297 | 82.68 | 0.32 | 0.138 | 83.35 | 0.36 | **0.095** | 84.13 | 0.16 | **0.029** |
| Core-set [37] | 79.69 | 0.82 | **0.043** | 81.41 | 0.45 | **0.096** | 82.25 | 0.39 | **0.021** | 83.11 | 0.38 | **0.032** | 83.73 | 0.31 | **0.006** |
| Multi-peak [27] | 80.36 | 0.4 | 0.225 | 81.48 | 0.53 | 0.125 | 82.63 | 0.23 | 0.119 | 83.29 | 0.2 | **0.036** | 84.3 | 0.44 | **0.063** |
| Learning Loss [51] | 79.58 | 0.39 | **0.002** | 81.39 | 0.34 | **0.071** | 82.31 | 0.42 | **0.038** | 83.31 | 0.25 | **0.029** | 84.2 | 0.53 | **0.098** |
| EGL++ (ours) | *80.49* | 0.45 | - | *81.91* | 0.27 | - | *83.03* | 0.43 | - | *83.91* | 0.51 | - | *84.68* | 0.36 | - |

## Uncertainty Quantification

- Single deterministic network
- Hyperparameter free (well … almost!)
- No modifications to existing architectures

# Thank you!

Megh Shukla
Computer Vision Research Engineer
Mercedes-Benz R&D India

megh.shukla@daimler.com
https://meghshukla.github.io