SAND2020-0579C

A Generalized Theory of Information

to Improve Algorithmic Learning and Predictions

Jed A. Duersch and Thomas A. Catanach

Joint Math Meeting, Denver, January 17, 2020

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Credible Prediction Uncertainty

- **What theoretical tools support analysis of generalizable machine learning?**
- How do we quantify **credible prediction uncertainty** on **abstract architectures**?
- Our investigation leads us to understand that **rational belief** serves a central role in the consistent measurement of **information as quantified change in belief**.

Published in Entropy 2020, doi.org/10.3390/e22010108:

Article

Generalizing Information to the Evolution of Rational Belief

Jed A. Duersch and Thomas A. Catanach

Sandia National Laboratories, Livermore, CA 94550, United States
Email: jaduers@sandia.gov and tacatan@sandia.gov

Problem: Mislabeled Data

- Suppose we wish to use a suitably trained ML model to identify mislabeled data.

Experimental example of states of belief:

Uninformed

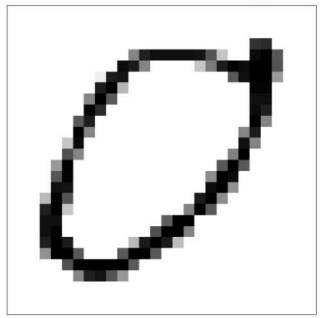
ML Predictions

Claimed Label

Correct Label

Digit	0	1	2	3	4	5	6	7	8	9
$q_0(y)$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
$q_1(y\|x)$	0.952	0.0	0.045	0.001	0.0	0.001	0.0	0.0	0.0	0.002
$r(y\|\check{y})$	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
$r(y\|\hat{y})$	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Actual image:



We would like an information measure to construct:

- Information gained by the prediction $q_0 \to q_1$.
- Residual information $q_1 \to r$.
- Ideally the sum would be a conserved quantity.

Problem: Mislabeled Data

Potential information measures based on Shannon's entropy:

- Kullback-Leibler divergence. $$D_{\mathrm{KL}}[q_1(y|x)\|q_0(y)] = \int dy\, q_1(y|x) \log\left(\frac{q_1(y|x)}{q_0(y)}\right)$$
- Lindley information. $$D_{\mathrm{L}}[q_1(y|x)\|q_0(y)] = \int dy\, q_1(y|x) \log(q_1(y|x)) - \int dy\, q_0(y) \log(q_0(y))$$

	Original Label			Corrected Label		
Information Type	**Prediction**	**Residual**	**Sum**	**Prediction**	**Residual**	**Sum**
Kullback-Leibler	3.02	10.44	13.46	3.02	0.07	3.09
Lindley	3.02	0.30	3.32	3.02	0.30	3.32

High KL residual information gives some useful indication

- KL does not conserve total information.
- Lindley information is conserved, but doesn't change.

Preview of proposed information measure:

	Original Label			Corrected Label		
Information Type	**Prediction**	**Residual**	**Sum**	**Prediction**	**Residual**	**Sum**
Proposed	-7.11	10.44	3.32	3.25	0.07	3.32

Reasonable Expectation (Cox, 1946)

The critical problem with the previous approaches is they do not respect the gravity of the role of rational belief in reasonable expectation.

Ensemble of independent realizations from $p(z)$: $Z = \{z_i \mid i \in [n]\}$

Ensemble mean of $f(z)$ approaches expectation: $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(z_i) = \mathbb{E}_{p(z)} f(z).$

What is true?

$$p(z) \mapsto \mathbb{E}_{p(z)} f(z)$$

What may be known?

$$r(z) \mapsto \mathbb{E}_{r(z)} f(z)$$

Digits of π example:

$$p(z) \equiv \delta(z - \pi)$$
$$\Pi = \{3.14159265\ldots, \ldots, 3.14159265\ldots\}$$

$$r(z) \equiv \mathcal{U}(3.14, 3.15)$$
$$\hat{\Pi} = \{3.148, 3.149, 3.141, 3.146, \ldots\}$$

What is Rational Belief?

- Cox (1946) uses **binary logic** to derive laws of probability as **an extended logic.**

Logic:

$$\sim\sim a = a, \quad (1)$$

$$a \cdot b = b \cdot a, \quad (2) \qquad a \vee b = b \vee a, \quad (2')$$

$$a \cdot a = a, \quad (3) \qquad a \vee a = a, \quad (3')$$

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c = a\ b \cdot c, \quad (4)$$

$$a \vee (b \vee c) = (a \vee b) \vee c = a \vee b \vee c, \quad (4')$$

$$\sim(a \cdot b) = \sim a \vee \sim b, \quad (5)$$

$$\sim(a \vee b) = \sim a \cdot \sim b, \quad (5')$$

$$a \cdot (a \vee b) = a, \quad (6) \qquad a \vee (a \cdot b) = a. \quad (6')$$

Extended logic:

- Probability is nonnegative.
- Impossibility has probability zero.
- Probability must be normalized to a constant.
- **Bayes' theorem conditions belief on evidence.**

$$p(a|b) = \frac{p(b|a)p(a)}{p(b)}$$

Dutch Book: An irrational agent accepts a table of bets that guarantee the agent will lose money.

Coin flip example:

Heads: 2:1 bet $2

Tails: 1:1 bet $3

} **Total bet $5. Guaranteed win $6.**

- **Rational belief requires a bet sequence to condition on previous results with Bayes' theorem (Skyrms, 1987).**

Information as Change in Belief

Postulates of **information as a rational measure of change in belief:**

1. **Information is an expectation over rational belief $r(z)$ that measures the change in belief from $q_0(z)$ to updated belief $q_1(z)$**

$$\mathbb{I}_{r(z)}[\, q_1(z) \,\|\, q_0(z) \,] = \int dz\, r(z) \underbrace{f(r(z), q_1(z), q_0(z))}_{\text{unspecified kernel}} .$$

2. **Information associated with independent processes is additive**

If $q_0(z,w) = q_0(z)q_0(w)$, $q_1(z,w) = q_1(z)q_1(w)$, and $r(z,w) = r(z)r(w)$ then

$$\mathbb{I}_{r(z,w)}[\, q_1(z,w) \,\|\, q_0(z,w) \,] = \mathbb{I}_{r(z)}[\, q_1(z) \,\|\, q_0(z) \,] + \mathbb{I}_{r(w)}[\, q_1(w) \,\|\, q_0(w) \,] .$$

3. **If belief does not change then information is zero** $\mathbb{I}_{r(z)}[\, q_0(z) \,\|\, q_0(z) \,] = 0.$

4. **Information gained from any proper hypothetical belief state to rational belief is nonnegative** $\mathbb{I}_{r(z)}[\, r(z) \,\|\, q_0(z) \,] \geq 0.$

Principal Result

Theorem 1. Information satisfying these postulates is computed as

$$\mathbb{I}_{r(z)}[\, q_1(z) \,\|\, q_0(z) \,] = \alpha \int dz\, r(z) \log\left(\frac{q_1(z)}{q_0(z)}\right).$$

for some $\alpha > 0$ corresponding to a choice of information units.

Information measures the change in belief from an **initial hypothesis $q_0(z)$** to **updated belief hypothesis $q_1(z)$** in the view of **rational belief $r(z)$**.

We recover **entropy, cross-entropy, KL divergence, Lindley information, mutual information,** and **other information measures** within this theory.

Information Pseudometrics

We may construct pseudometrics that **measure distance between states of belief $q_0(z)$ and $q_1(z)$** with the view of expectation $r(z)$,

by taking weighted-L^P norms of information density

$$\mathbb{L}^p_{r(z)}\left[q_1(z)\|q_0(z)\right] = \left(\int dz\, r(z) \left|\log\left(\frac{q_1(z)}{q_0(z)}\right)\right|^p\right)^{1/p} \quad \text{for any } p > 1.$$

These pseudometrics are **weak discriminator functions**; they identify a weakened form of equivalence in belief states subject to rational belief.

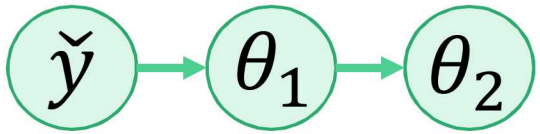
Selected Corollaries

Corollary 1 shows how we can measure information over **arbitrary correlations between multiple variables** in each state of belief as

$$\mathbb{I}_{r(z_1,z_2)}\left[q_1(z_1,z_2)\|q_0(z_1,z_2)\right] = \mathbb{I}_{r(z_1)}\left[q_1(z_1)\|q_0(z_1)\right] + \mathbb{E}_{r(z_1)}\mathbb{I}_{r(z_2|z_1)}\left[q_1(z_2|z_1)\|q_0(z_2|z_1)\right].$$

We show how this and other corollaries help us derive information bounds such as **Corollary 14, Monotonically decreasing local inference information:**

$$\mathbb{I}_{p(\theta_2|\check{y})}\left[p(\theta_2|\check{y})\|p(\theta_2)\right] \leq \mathbb{I}_{p(\theta_1|\check{y})}\left[p(\theta_1|\check{y})\|p(\theta_1)\right].$$

where 

Note that using training data to infer model belief, which then yields predictions $\check{y} \mapsto \theta \mapsto \check{y}$ is a locally conditioned sequence.

Proper Utility (Bernardo, 1981)

- Bernardo considers utility functions (objectives for optimization) expressible as **expectations over rational belief**.

$$\mathcal{U}\left[r(z),\, q(z)\right] = \int dz\, r(z) f\left(r(z), q(z)\right).$$

- For a utility function to be proper, **rational belief must be the unique optimizer**.

$$q^*(z) \equiv r(z) \equiv \arg\max_{q(z)} \mathcal{U}\left[r(z),\, q(z)\right]$$

- **Corollaries 10 and 11, information is a proper utility function**

$$\mathbb{I}_{r(z)}\left[\, q(z) \,\|\, q_0(z) \,\right].$$

Information in Training Labels

A second additivity property follows; **Corollary 2 – information gained over a sequence of belief updates is additive** within the same view:

$$\mathbb{I}_{r(z)}[\, q_2(z) \,\|\, q_0(z) \,] = \mathbb{I}_{r(z)}[\, q_2(z) \,\|\, q_1(z) \,] + \mathbb{I}_{r(z)}[\, q_1(z) \,\|\, q_0(z) \,] \,.$$

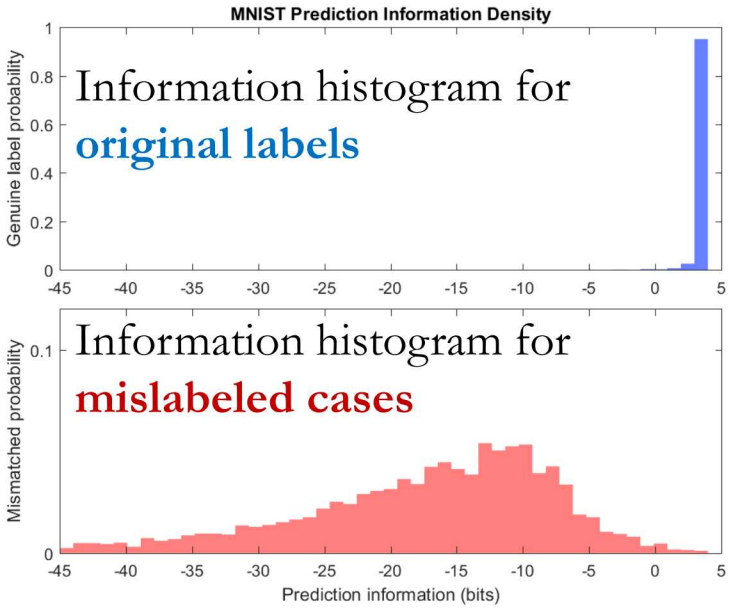
Predictive label information: $\mathbb{I}_{r(y|\check{y})}[\, q_1(y|x) \,\|\, q_0(y) \,] \,.$

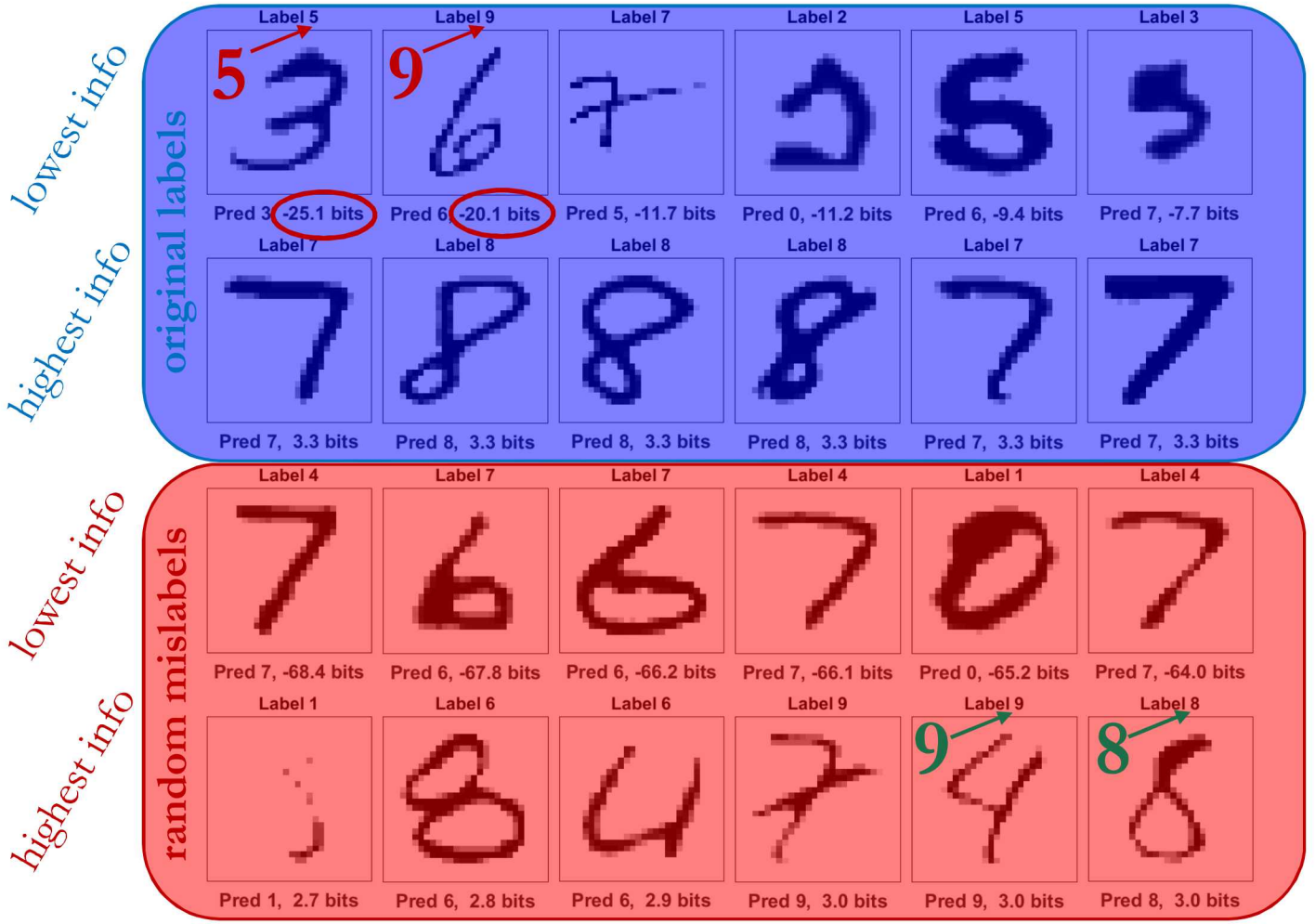
Residual label information: $\mathbb{I}_{r(y|\check{y})}[\, r(y|\check{y}) \,\|\, q_1(y|x) \,] \,.$

Total label information: $\mathbb{I}_{r(y|\check{y})}[\, r(y|\check{y}) \,\|\, q_0(y) \,] \,.$

Detecting Mislabeled Cases

We compute prediction information in a trained neural network and examine prediction information on previously unseen cases. Half are **randomly mislabeled**.





Poorly Trained Model

In this experiment, the ML model was **trained with some incorrect labels.**

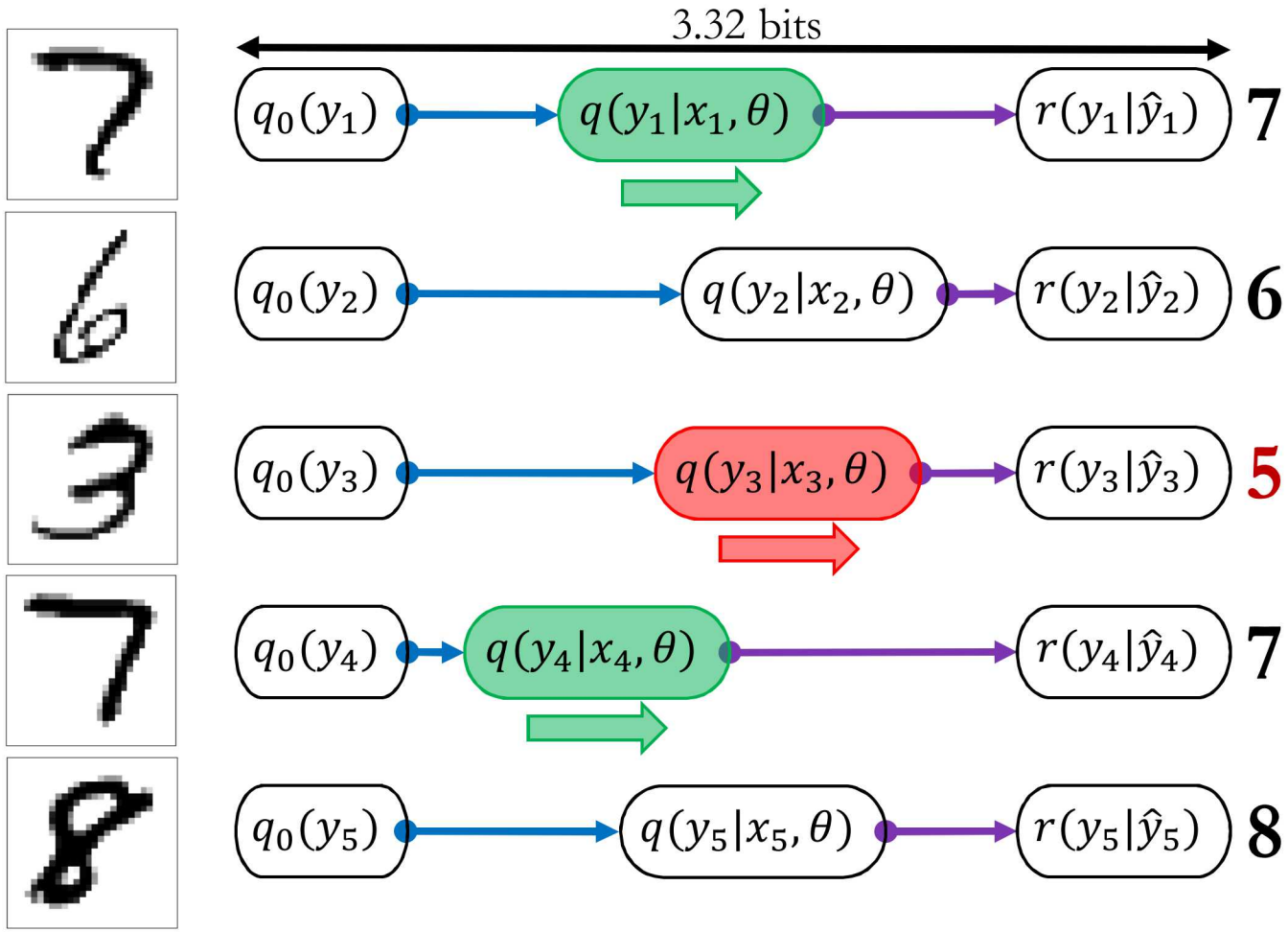
- **Cross-validation optimum has learned to memorize incorrect labels.**
- Again, neither KL nor Lindley constructions of prediction information give any indication of the poor performance of this model.

Digit	0	1	2	3	4	5	6	7	8	9
$q_0(y)$	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
$q_1(y\|x)$	0.03	0.00	0.61	0.04	0.00	0.00	0.02	0.00	0.30	0.00
$r(y\|\check{y})$	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r(y\|\hat{y})$	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00

	Original Labels			Corrected Labels		
Information Type	**Prediction**	**Residual**	**Sum**	**Prediction**	**Residual**	**Sum**
Kullback-Leibler	1.87	0.72	2.59	1.87	4.78	6.65
Lindley	1.87	1.45	3.32	1.87	1.45	3.32
Proposed	2.60	0.72	3.32	-1.46	4.78	3.32

Learning and Memorization

The information in each training example has a **predicted component** and a **residual component**.



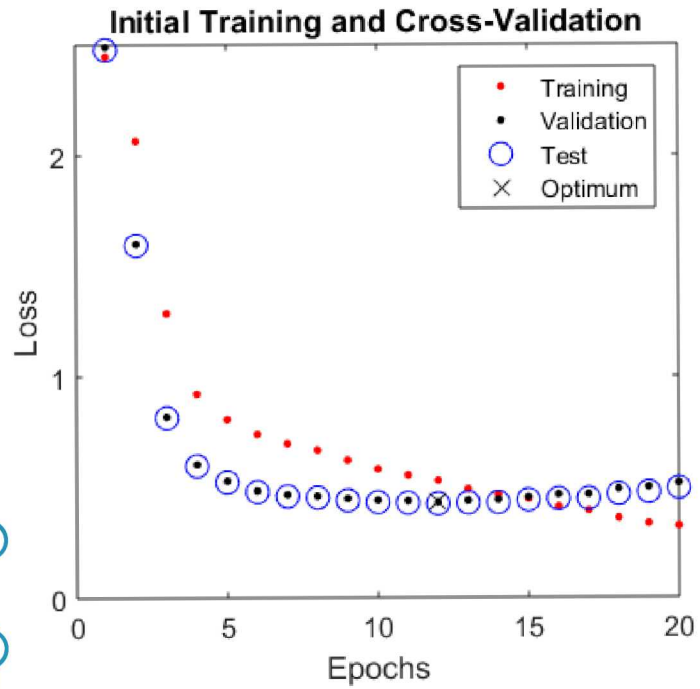
Pattern learning – we alter the model and **multiple cases benefit**.

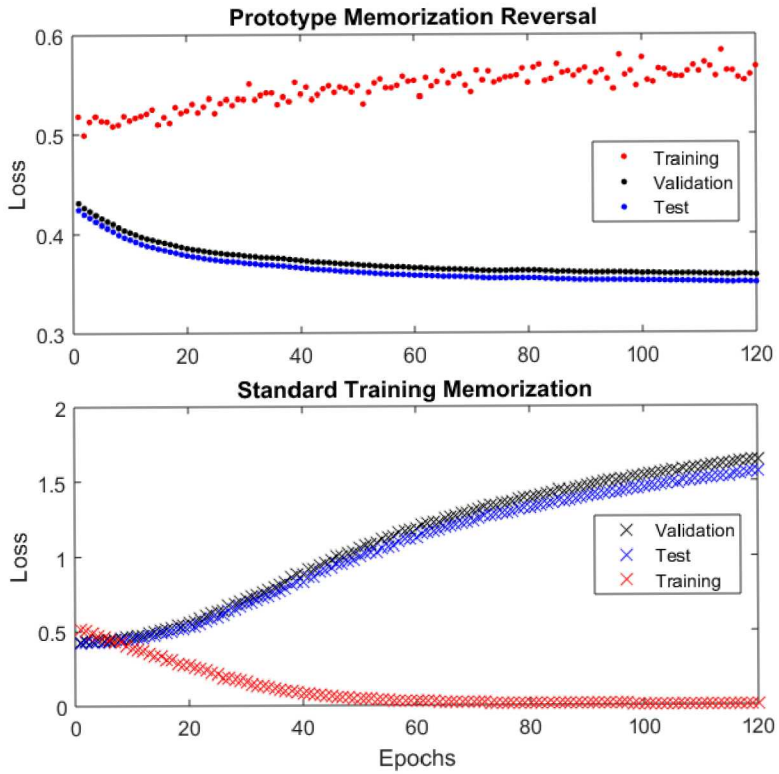
Memorization – we alter the model and **only a single case benefits.**

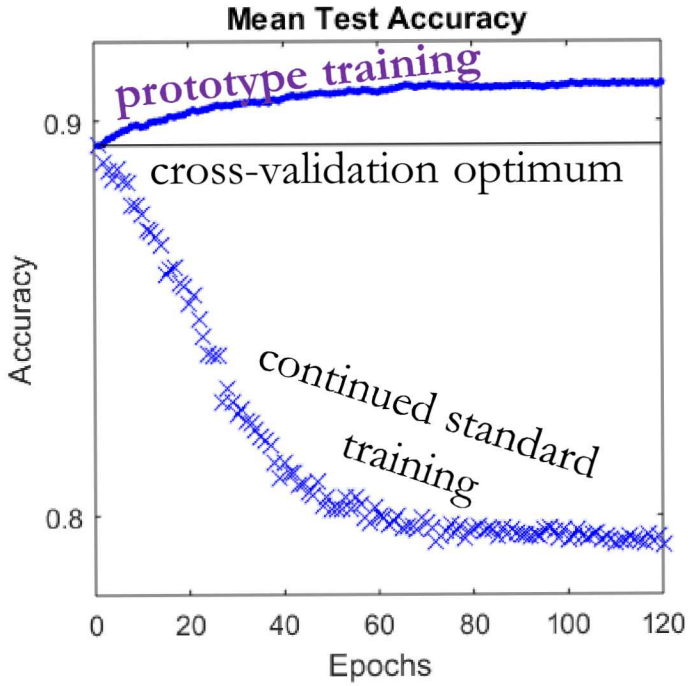
Memorization Reversal

We can demonstrate **memorization reversal** on partially mislabeled MNIST data from the cross-validation optimum. **The result surpasses the best model obtainable from standard techniques.**

$$\mathrm{Cov}(g) \approx U\left[\mathrm{diag}\,\sigma\right]^2 U^T \qquad \hat{g} = U \max_{\alpha \geq 1}(0, 1 - \alpha\sigma) \circledast U^T \bar{g}$$





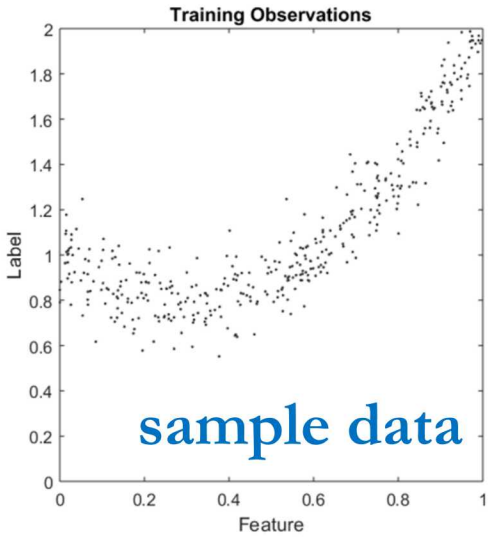


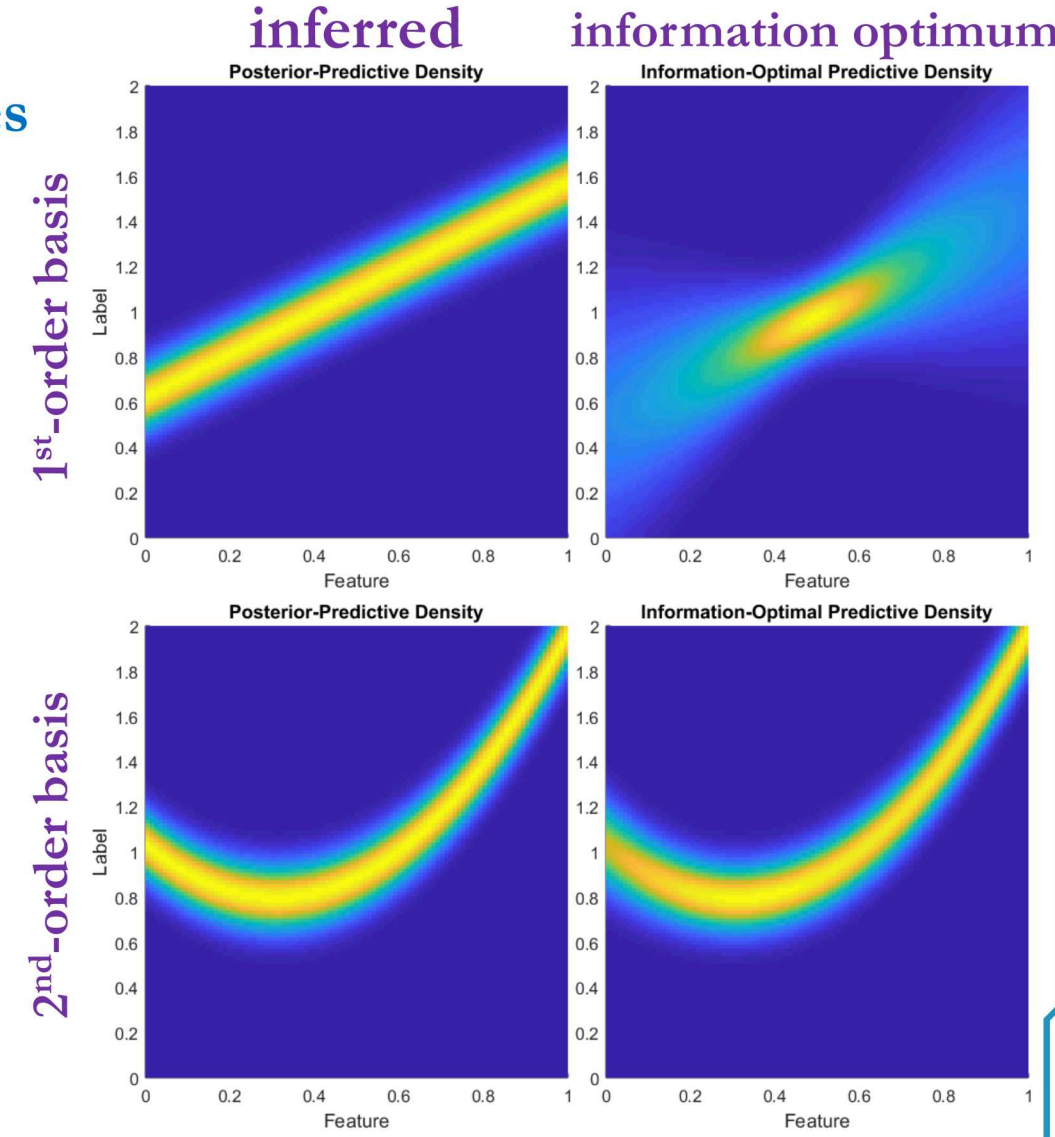
Information Optimality

Inference on **under-expressive model families** can yield **poor predictions**.

We can **improve prediction uncertainty** credibility by **optimizing information** over distributions of models from the family.

KL residual information is not suitable for optimization in this case because it **is infinite**.





Future Work

- Efficiently **controlling model complexity** during model training.
- Evaluating and controlling the **influence of individual data points** on the trained model and resulting predictions.
- Robust **anomaly detection**.
- Understanding the influence of **architectural design** on predictability.

Published in Entropy 2020, doi.org/10.3390/e22010108:

Article

Generalizing Information to the Evolution of Rational Belief

Jed A. Duersch and Thomas A. Catanach

Sandia National Laboratories, Livermore, CA 94550, United States
Email: jaduers@sandia.gov and tacatan@sandia.gov

Thank you!