# Uniform Sampling over Episode Difficulty

Sébastien M. R. Arnold*,[1], Guneet S. Dhillon*,[2], Avinash Ravichandran[3], Stefano Soatto[3,4]
seb.arnold@usc.edu, guneet.dhillon@stats.ox.ac.uk, ravinash@amazon.com, soattos@amazon.com
[1]University of Southern California, [2]University of Oxford, [3]Amazon Web Services, [4]University of California, Los Angeles
*Equal Contribution, done at AWS

## Summary

**We ask**

- *How to best sample episodes for few-shot classification?*

**TL;DR**

- Sampling **uniformly over episode difficulty** outperforms curriculum learning, easy / hard mining, and the standard baseline.

**Our Contributions**

- **Extensive study of episode difficulty**, which shows that difficulty is approx. normally distributed, and that (relative) episode difficulty is invariant to model architecture, training algorithm, or model parameters.
- Propose **importance sampling over episode difficulty** to implement 5 sampling schemes for episodic training.
- Show that **sampling episodes uniformly over difficulty outperforms other schemes**, on many few-shot and cross-domain benchmarks.

## Background

**Classification episode** $\tau$ consist of

- support set $\tau_S$: few $(x, y)$ samples used to (quickly) solve classification task encoded by episode.
- query set $\tau_Q$: samples used to evaluate the how well the episode was solved.

**Episodic training** denotes maximizing the likelihood of all possible episodes under model parameters $\theta$:

$$\max_{\theta} \mathbb{E}_{\tau}\left[\ell\left(\tau_Q \mid \theta, \tau_S\right)\right]$$

**Episode difficulty**, in our case, is the negative log-likelihood of episode $\tau$:

$$\Omega_\theta(\tau) = -\log \ell\left(\tau_Q \mid \theta, \tau_S\right)$$

### Importance Sampling for Episodic Training

**Importance Sampling (IS)** reweights a sample $x$ from proposal distribution $q(x)$ with weight $\frac{p(x)}{q(x)}$ so it comes from target distribution $p(x)$.

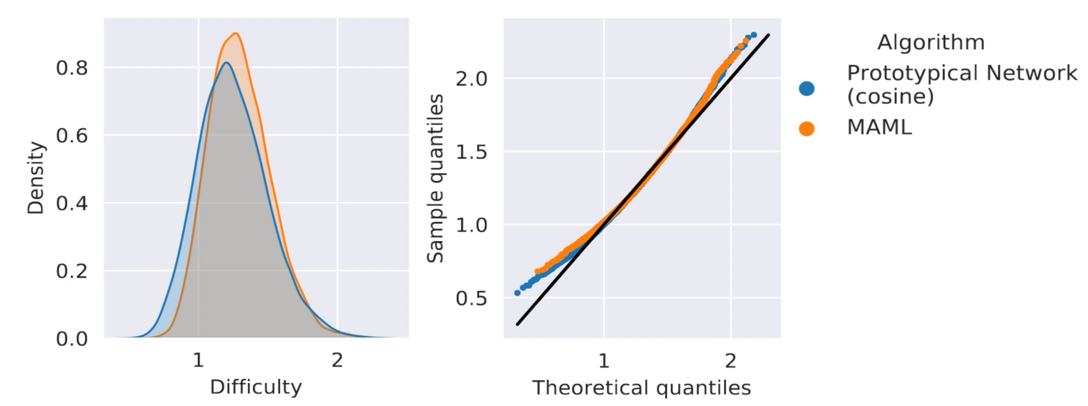
**Importance Sampling for Episodes** uses IS to sample episodes from target difficulty distribution $q(\tau)$ from proposal difficulty distribution $p(\tau)$:

$$\mathbb{E}_{\tau\sim p}[\Omega_\theta(\tau)] = \mathbb{E}_{\tau\sim q}\left[\frac{p(\tau)}{q(\tau)}\,\Omega_\theta(\tau)\right]$$

Use **Effective Sample Size** to adjust for small mini-batches.

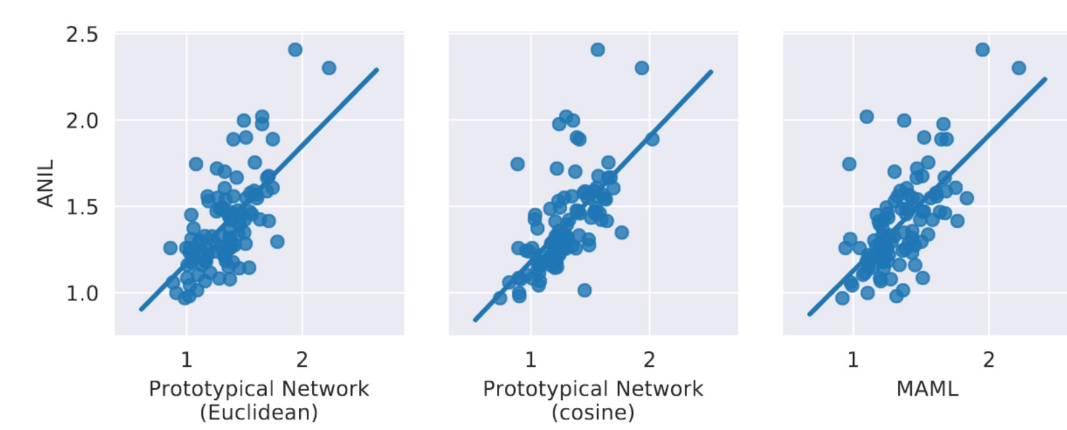
## Understanding Episode Difficulty

**Episode difficulty is ~ normally distributed.**



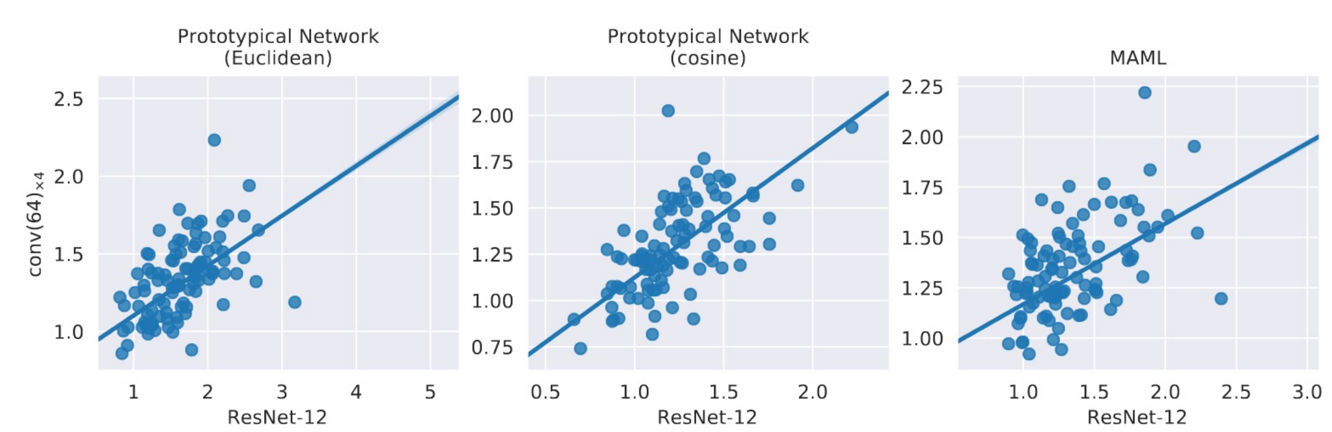

Density and Q-Q plots of episode difficulty for ProtoNet and MAML on 5-ways 1-shot Mini-ImageNet episodes. The bell curve and (almost) linear relationship suggest difficulty is approximately normally distributed.

**Episode transfers across training algorithms.**



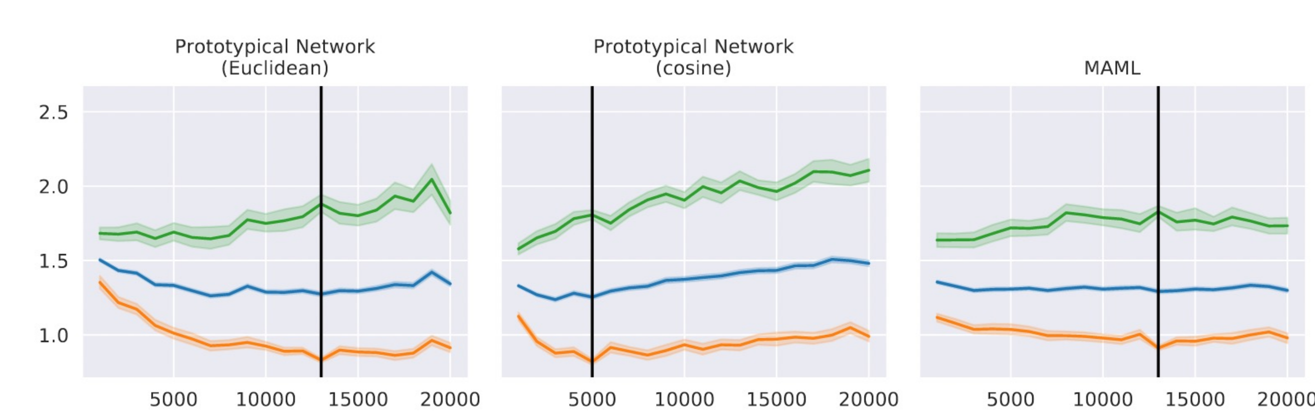

Correlation of episode difficulty for ANIL and ProtoNet (Euclidean), ProtoNet (Cosine), or MAML on 10k 5-ways 1-shot Mini-ImageNet episodes. Spearman's correlation > 0.65 for all, showing difficulty transfers across algorithms.

**Episode transfers across network architectures.**



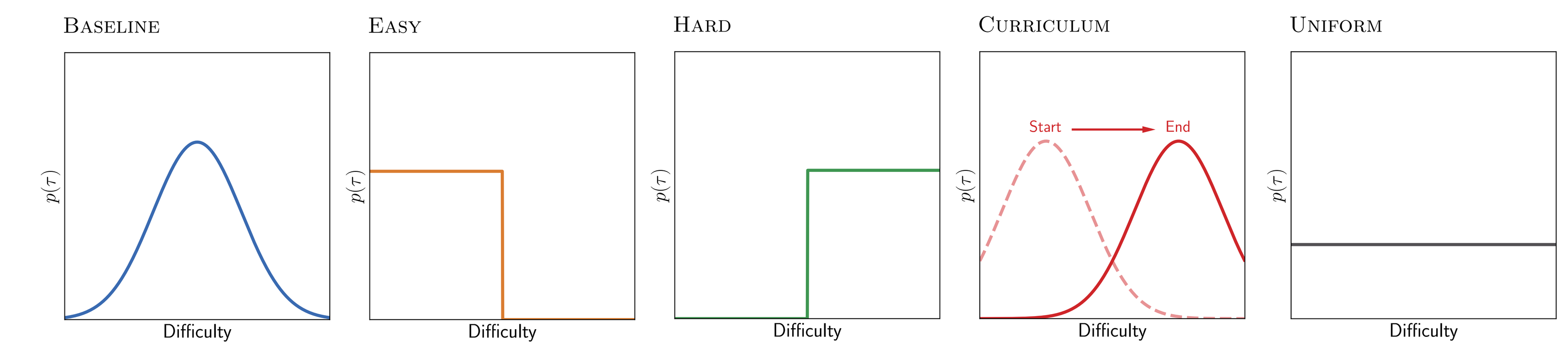

Correlation of episode difficulty for a 4-layer CNN and a 12-layer ResNet (trained with ProtoNet or MAML) on 10k 5-ways 1-shot Mini-ImageNet episodes. Spearman's correlation > 0.55 for all, showing difficulty transfers across architectures.

**Episode transfers across model parameters.**




Tracking the average difficulty for the 50 easier and 50 harder Mini-ImageNet episodes during training. Easy episodes stay easy; hard episodes stay hard.

## Episode Sampling Schemes

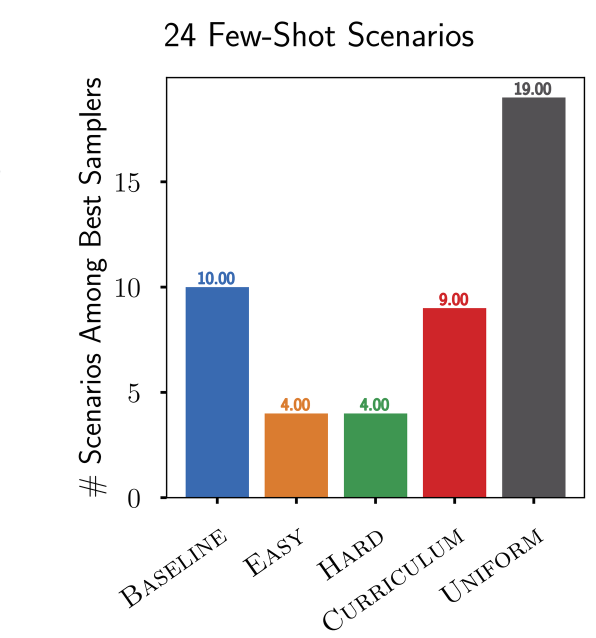



With current difficulty mean $\mu$ and difficulty variance $\sigma^2$:

- Baseline: $\tau \sim \mathcal{N}(\mu, \sigma^2)$
- Easy: $\tau \sim \mathcal{U}(\mu - 2.58\,\sigma,\ \mu)$
- Hard: $\tau \sim \mathcal{U}(\mu,\ \mu + 2.58\,\sigma)$
- Curriculum: $\tau \sim \mathcal{N}(\mu_t, \sigma^2)$ with $\mu_t = \mu + 2.58(2t - 1)\sigma$
- Uniform: $\tau \sim \mathcal{U}(\mu - 2.58\,\sigma,\ \mu + 2.58\,\sigma)$

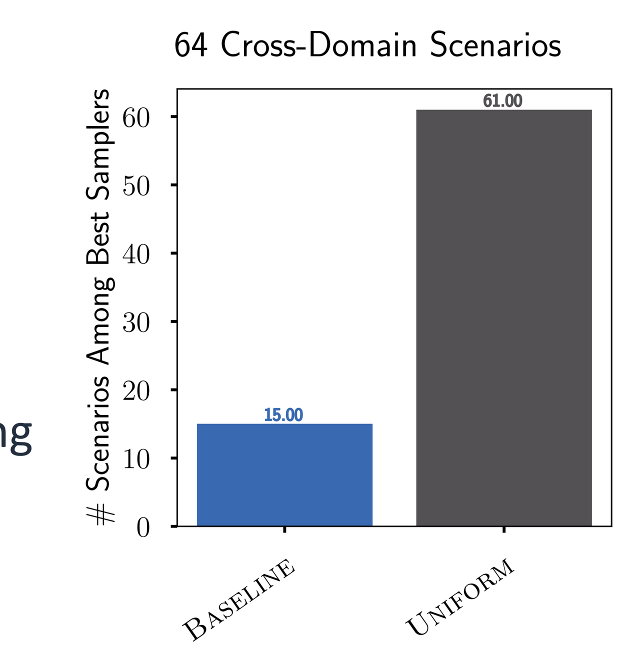
## Improving Few-Shot Learning

- Compare sampling schemes across 24 scenarios (varying architectures, algorithms, datasets, settings).
- Tally how often a scheme is among the best performing.
- Uniform dominates; Baseline & Curriculum second best.




## Improving Cross-Domain Transfer

- Zero in on Uniform vs Baseline for cross-domain transfer.
- Pretrain models on Mini-ImageNet; evaluate accuracy when transferred to 4 unseen datasets. Total: 64 scenarios.
- Uniform is best on 49/64 scenarios, among best on 61/64.




## Learn More

In the paper: full tables, improving SOTA with FEAT, online vs offline samplers, and more...

Available at http://sebarnold.net/projects/eis