Active Learning in Bayesian Neural Networks with Balanced Entropy Learning Principle

Jae Oh Woo*

Samsung SDS America

*Opinions are my own

ICLR
International Conference On Learning Representations

SAMSUNG SDS

Active Learning (simplified)

- ❑ Unlabeled data pool

- ❑ Select acquiring data points from unlabeled data pool

- ❑ Annotate ground truth labels by human

- ❑ Train neural network with labeled dataset

Active Learning (simplified)

- Unlabeled data pool

- Select acquiring data points from unlabeled data pool

- Annotate ground truth labels by human

- Train neural network with labeled dataset

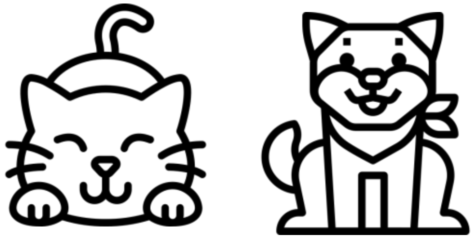
SAMSUNG SDS

Active Learning (simplified)

- ❑ Unlabeled data pool

- ❑ Select acquiring data points from unlabeled data pool

- ❑ Annotate ground truth labels by human



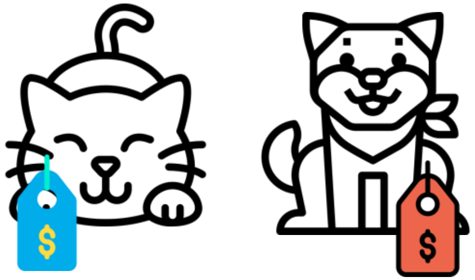
- ❑ Train neural network with labeled dataset

Active Learning (simplified)

- ❑ Unlabeled data pool

- ❑ Select acquiring data points from unlabeled data pool

- ❑ Annotate ground truth labels by human



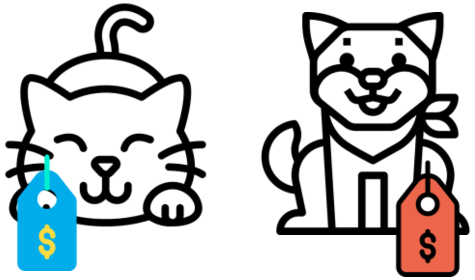
- ❑ Train neural network with labeled dataset

Active Learning (simplified)

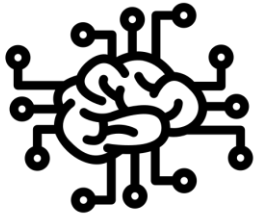
- ❑ Unlabeled data pool

- ❑ Select acquiring data points from unlabeled data pool

- ❑ Annotate ground truth labels by human



- ❑ Train neural network with labeled dataset

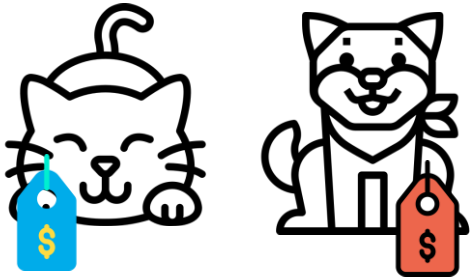


Active Learning (simplified)

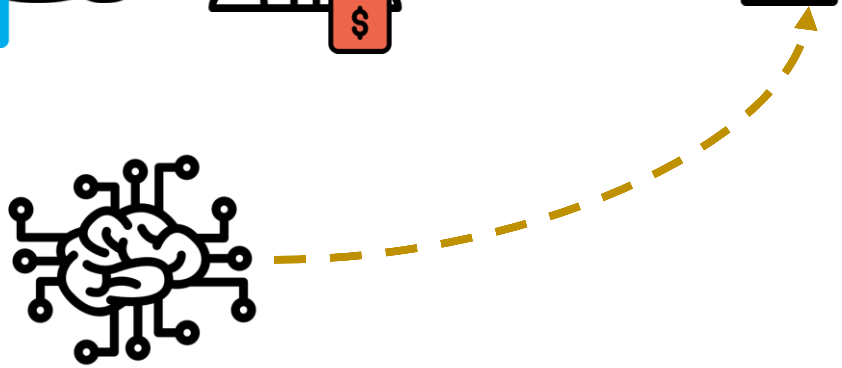
- ❑ Unlabeled data pool

- ❑ Select acquiring data points from unlabeled data pool

- ❑ Annotate ground truth labels by human



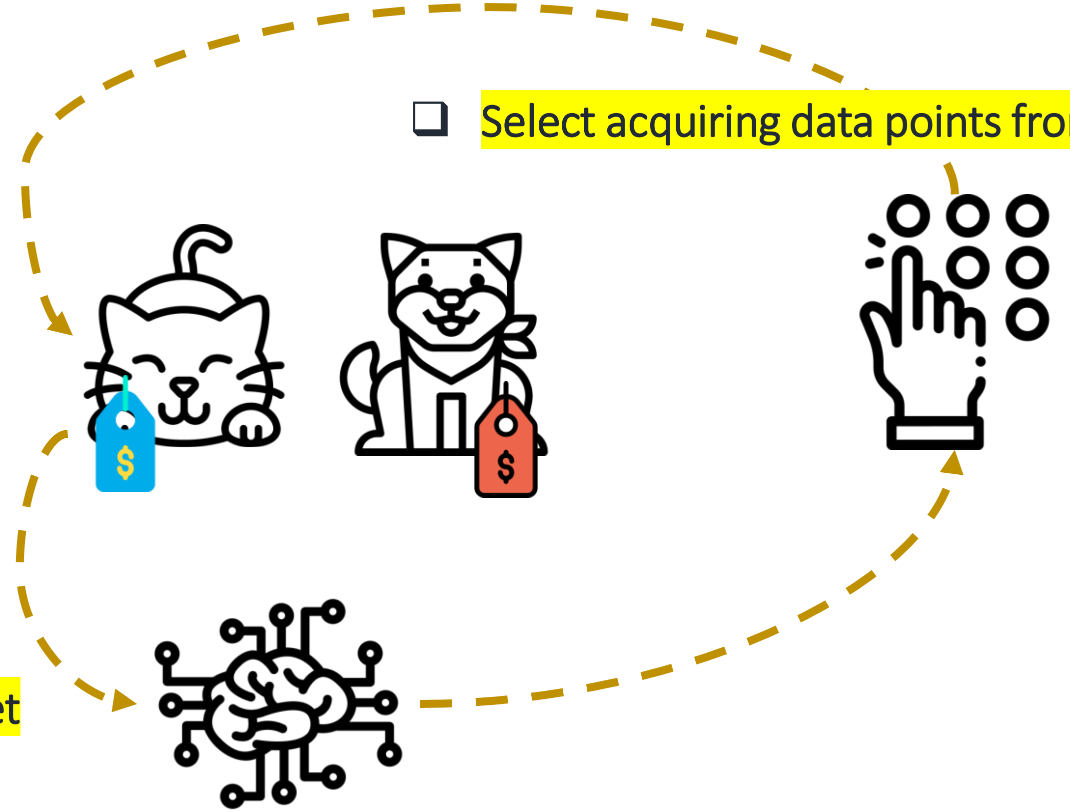
- ❑ Train neural network with labeled dataset



Active Learning (simplified)

- ❑ Unlabeled data pool

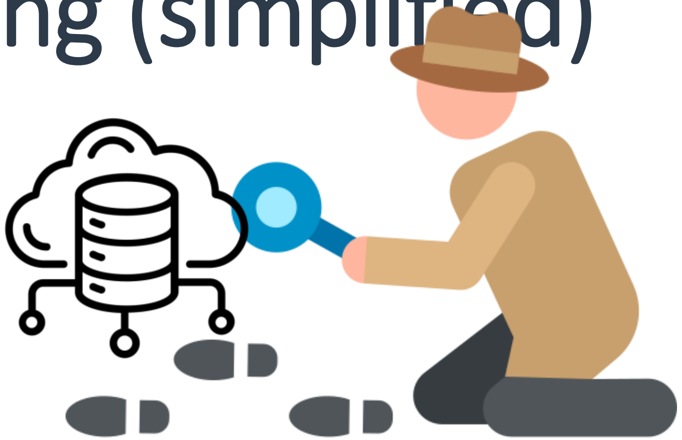
- ❑ Select acquiring data points from unlabeled data pool



- ❑ Annotate ground truth labels by human
- ❑ Train neural network with labeled dataset

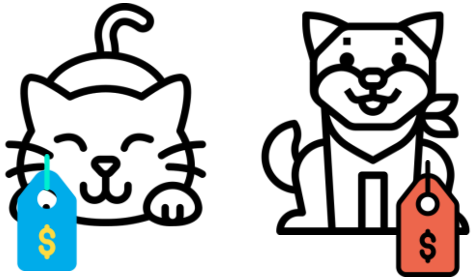
Active Learning (simplified)

❑ Unlabeled data pool



✓ Select acquiring data points from unlabeled data pool

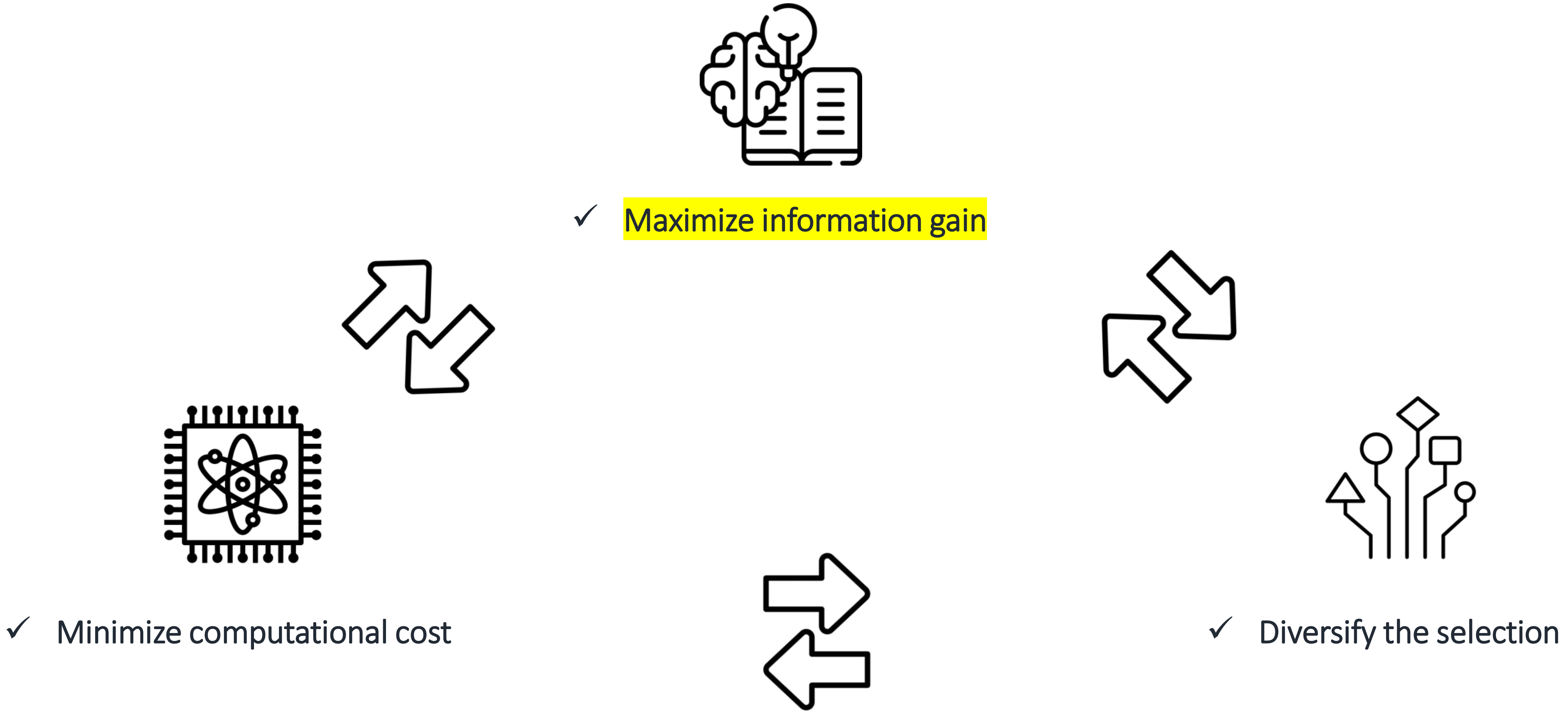
❑ Annotate ground truth labels by human



❑ Train neural network with labeled dataset

SAMSUNG SDS

Tradeoffs



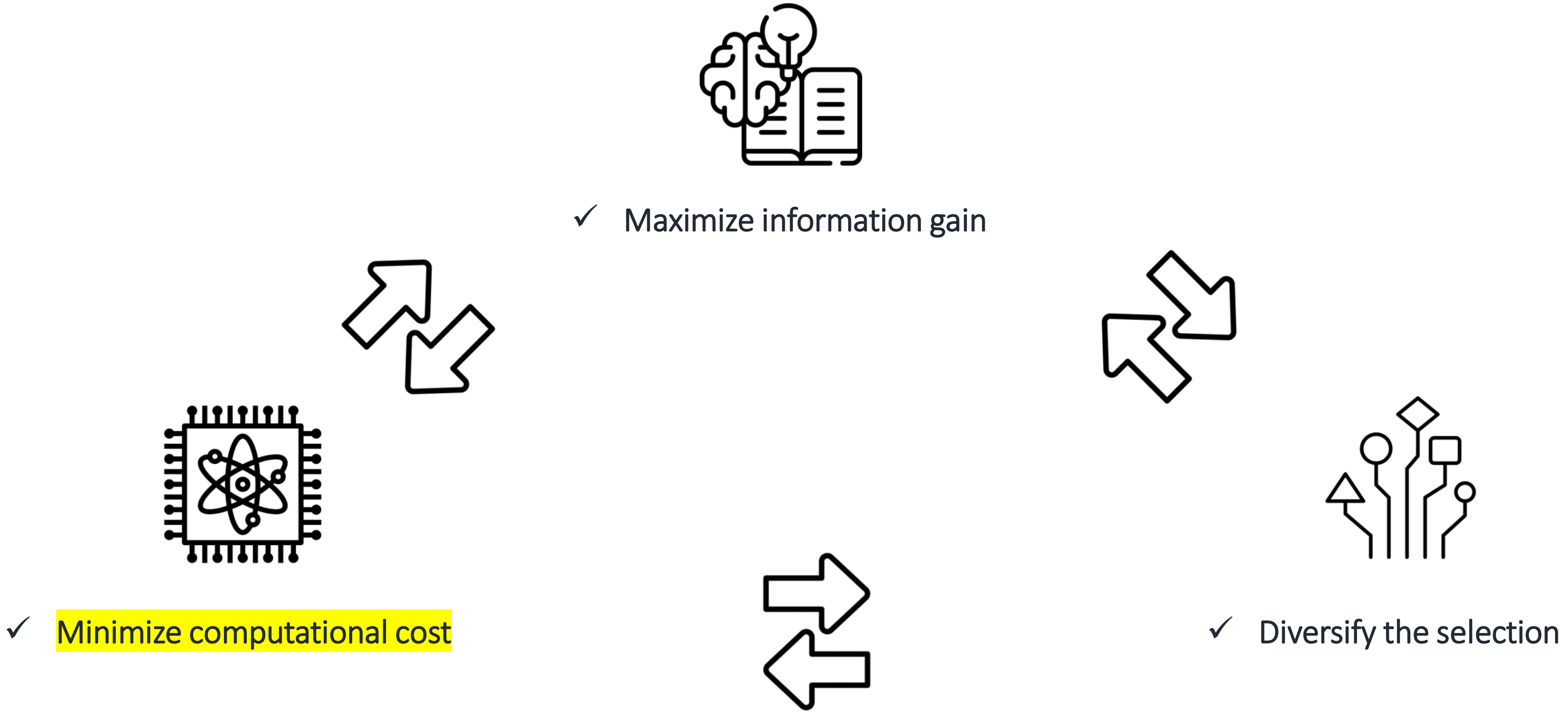
✓ Maximize information gain
✓ Minimize computational cost
✓ Diversify the selection

Tradeoffs

- ✓ Maximize information gain
- ✓ Minimize computational cost
- ✓ Diversify the selection

SAMSUNG SDS

Tradeoffs


✓ Maximize information gain
✓ Minimize computational cost
✓ Diversify the selection

Tradeoffs

- BALD – Mutual Information
- Entropy
- MeanSD
- Variational Ratio
- …

✓ Maximize information gain

✓ Minimize computational cost

✓ Diversify the selection

[1] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In International Conference on Machine Learning, pp. 1183–1192. PMLR, 2017.
[2] Claude E Shannon. A mathematical theory of communication. The Bell system technical journal, 27(3):379–423, 1948.
[3] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. Journal of artificial intelligence research, 4:129–145, 1996.
[4] L.C. Freeman. Elementary Applied Statistics: For Students in Behavioral Science. For Students in Behavioral Science. Wiley, 1965.

Tradeoffs

- BADGE
- BatchBALD
- CoreSet
- ...

✓ Maximize information gain

✓ Minimize computational cost

✓ Diversify the selection

[1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. International Conference on Learning Representations, 2020.
[2] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. Advances in Neural Information Processing Systems 2019.
[3] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In International Conference on Learning Representations, 2018.

Tradeoffs

✓ Maximize information gain

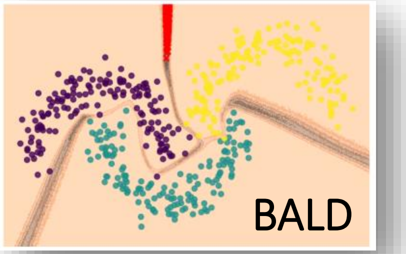
○ PowerBALD

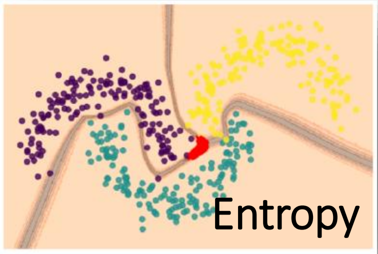
✓ Minimize computational cost

✓ Diversify the selection

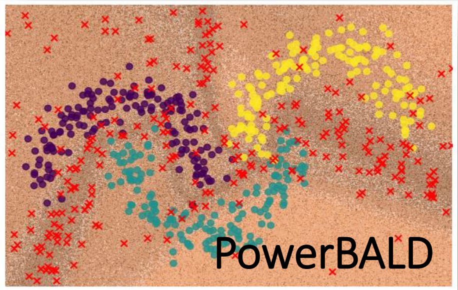
[1] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. International Conference on Learning Representations, 2021.
[2] Andreas Kirsch, Sebastian Farquhar, and Yarin Gal. A simple baseline for batch active learning with stochastic acquisition functions. arXiv preprint arXiv:2106.12059, 2021.

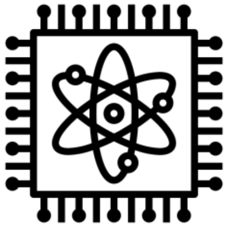
Tradeoffs



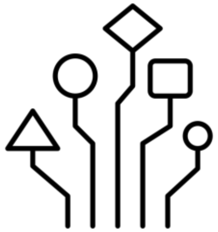


✓ Maximize information gain





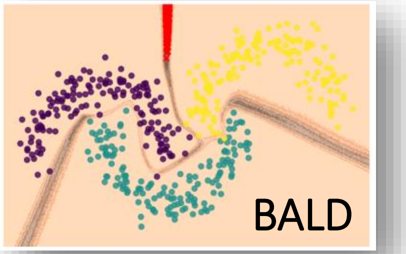
✓ Minimize computational cost

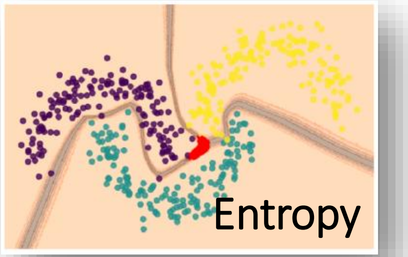


✓ Diversify the selection

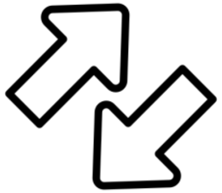
SAMSUNG SDS

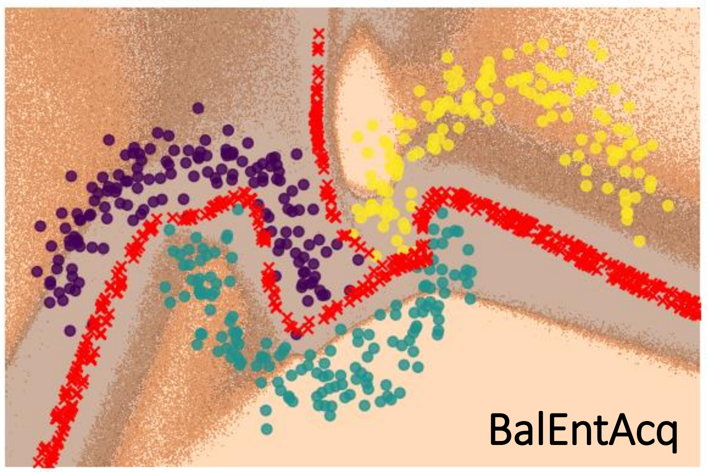
Tradeoffs

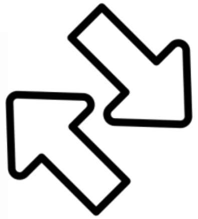


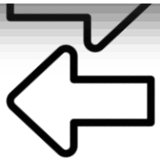


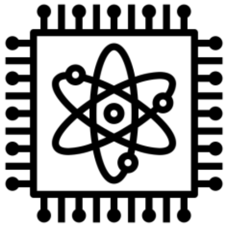
✓ Maximize information gain



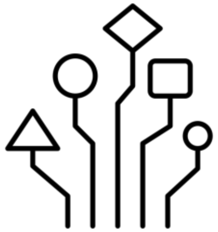








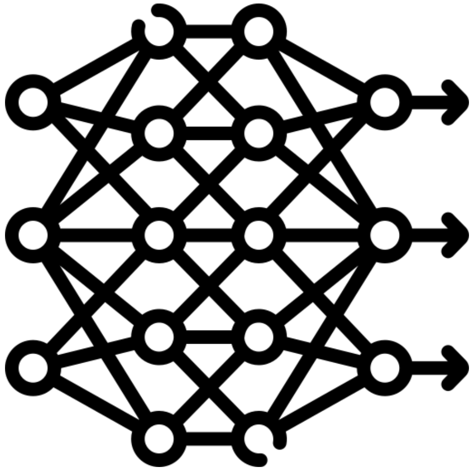
✓ Minimize computational cost



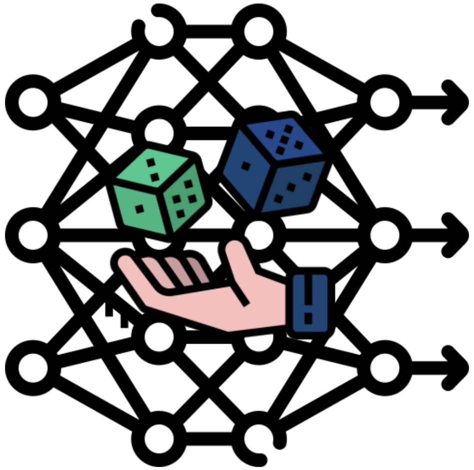
✓ Diversify the selection

SAMSUNG SDS

Bayesian Neural Networks



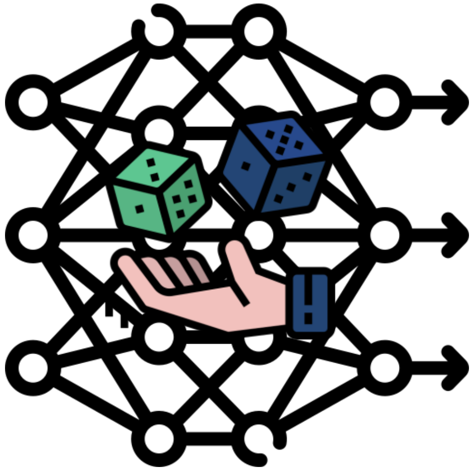
Bayesian Neural Networks



- ✓ Apply dropouts
- ✓ Laplace Approximation
- ✓ …

[1] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In international conference on machine learning, pp. 1050–1059. PMLR, 2016
[2] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In International Conference on Machine Learning, pp. 5436–5446. PMLR, 2020
[3] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. Advances in Neural Information Processing Systems, 34, 2021.

Bayesian Neural Networks

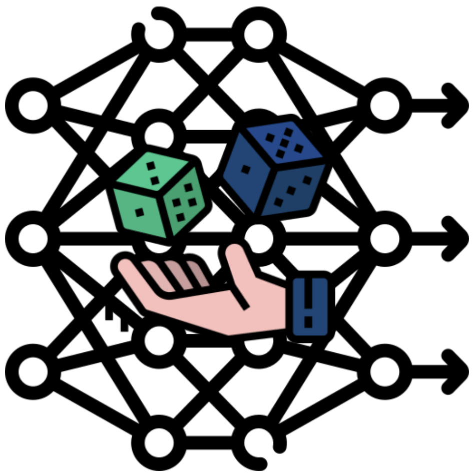


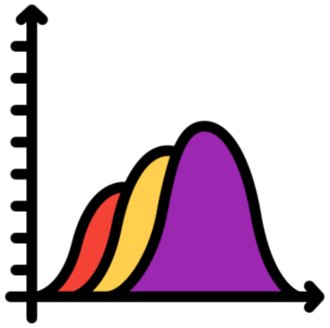
- ✓ Apply dropouts
- ✓ Laplace Approximation
- ✓ …

[1] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In international conference on machine learning, pp. 1050–1059. PMLR, 2016
[2] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In International Conference on Machine Learning, pp. 5436–5446. PMLR, 2020
[3] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. Advances in Neural Information Processing Systems, 34, 2021.

SAMSUNG SDS

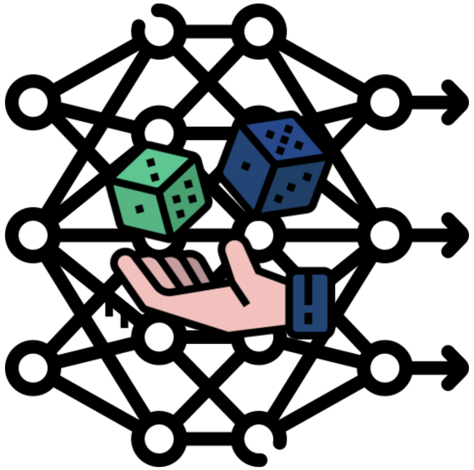
Bayesian Neural Networks

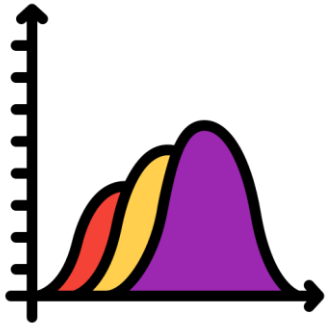




✓ Apply dropouts

Bayesian Neural Networks

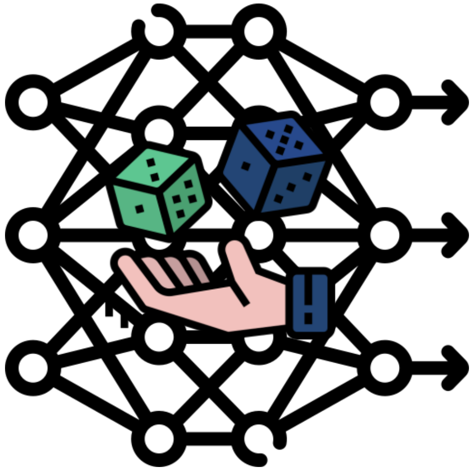


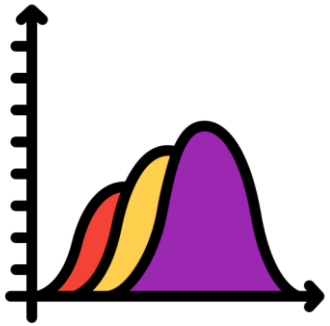


softmax

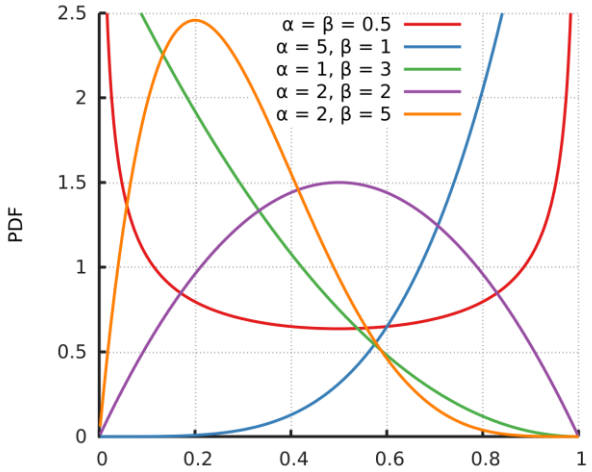
✓ Apply dropouts

Bayesian Neural Networks + Beta Approximation





≈
softmax

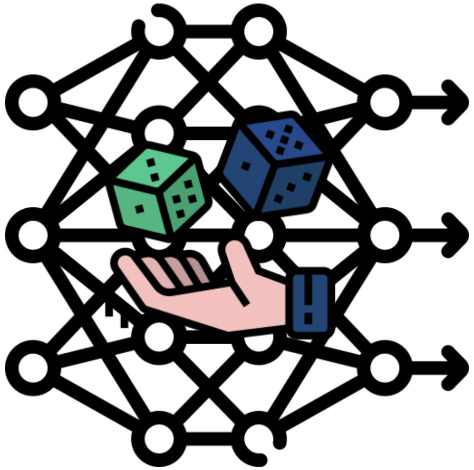


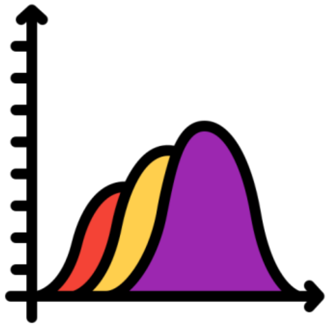
- ✓ Apply dropouts
- ✓ Beta distribution for each marginal

[1] https://en.wikipedia.org/wiki/Beta_distribution

SAMSUNG SDS

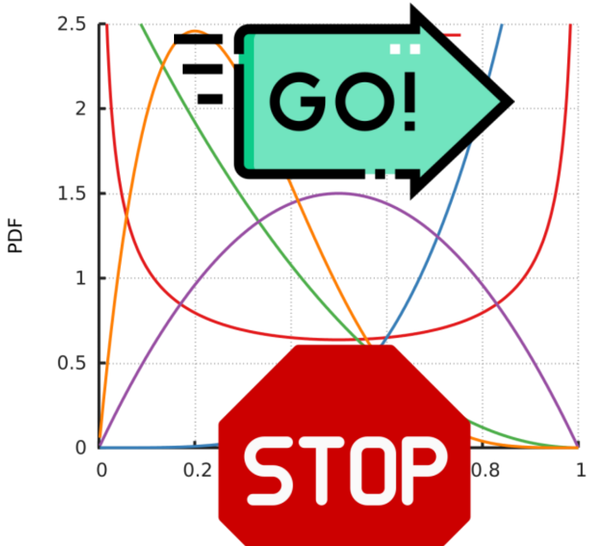
Bayesian Neural Networks + Beta Approximation





≈

softmax

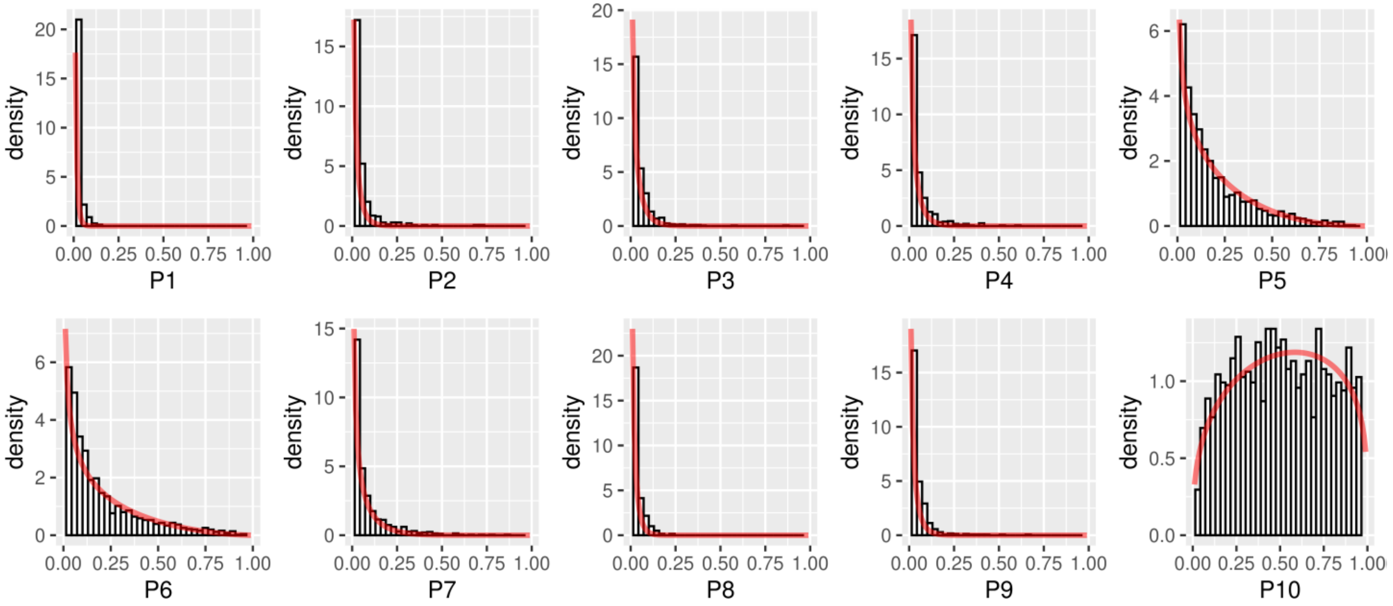


✓ Apply dropouts

✓ Beta distribut[illegible]each marginal

[1] https://en.wikipedia.org/wiki/Beta_distribution

Beta Distribution Approximation



✓ MNIST

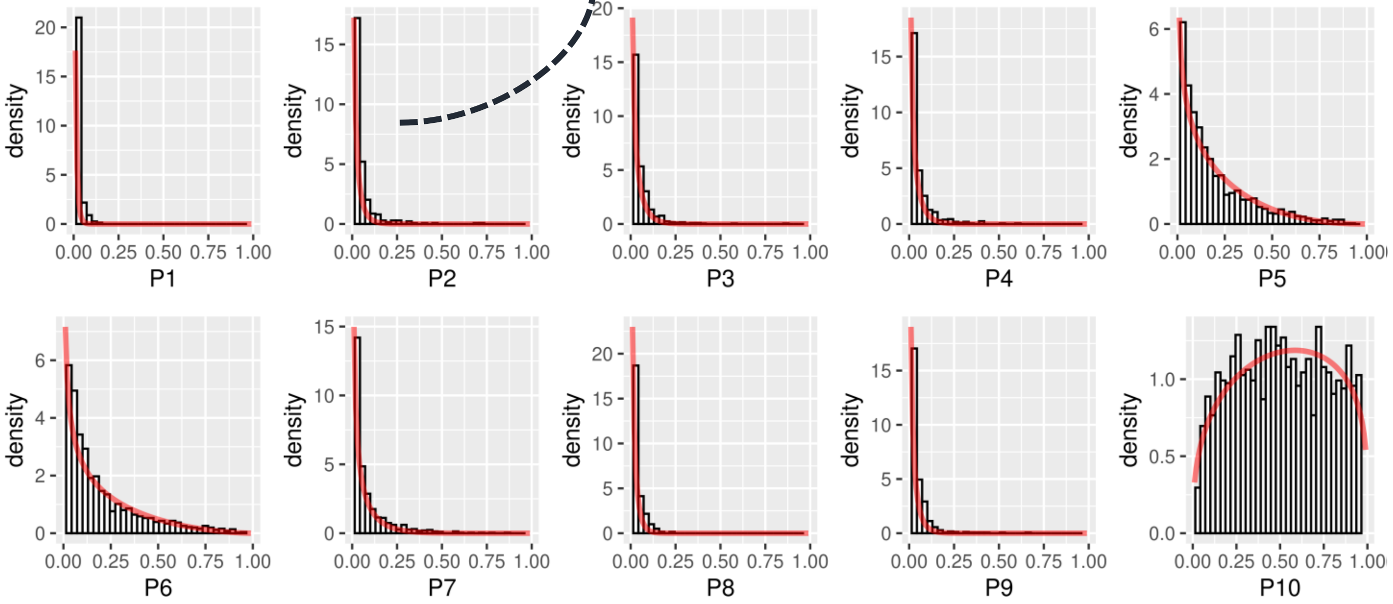
SAMSUNG SDS

Beta Distribution Approximation

✓ MNIST

Match mean and variance

$$\begin{cases} P_i \sim \mathrm{Beta}(\alpha_i, \beta_i), \\ P_i^+ \sim \mathrm{Beta}(\alpha_i + 1, \beta_i) \end{cases}$$

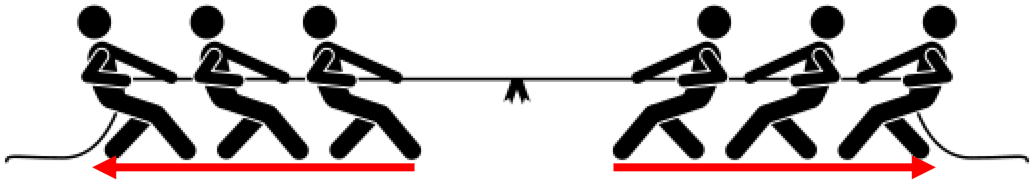


Balanced Entropy

$$-\infty \leq \boxed{\mathbf{BalEnt}[\mathbf{x}] := \frac{\sum \mathrm{E} P_i h(P_i^+) + H(Y)}{H(Y) + \log 2}} \leq 1, \qquad Y = \text{Class Label}$$

$\sum \mathrm{E} P_i h(P_i^+) \leq 0$

✓ Continuous differential entropy

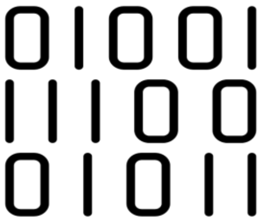


✓ Model confidence

✓ Label uncertainty

$H(Y) \geq 0$

✓ Discrete Shannon entropy

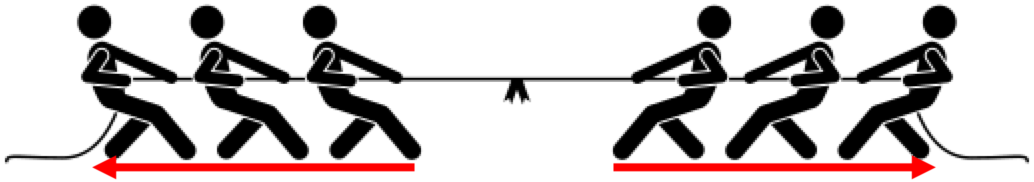


Balanced Entropy

$$-\infty \leq \mathbf{BalEnt[x]} := \frac{\sum \mathrm{E} P_i h(P_i^+) + H(Y)}{H(Y) + \log 2} \leq 1, \qquad Y = \text{Class Label}$$

$\sum \mathrm{E} P_i h(P_i^+) \leq 0$

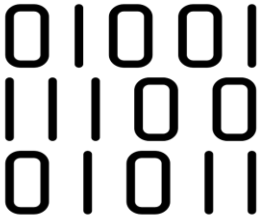
- ✓ Continuous differential entropy



- ✓ Model confidence
- ✓ Label uncertainty

$H(Y) \geq 0$

- ✓ Discrete Shannon entropy

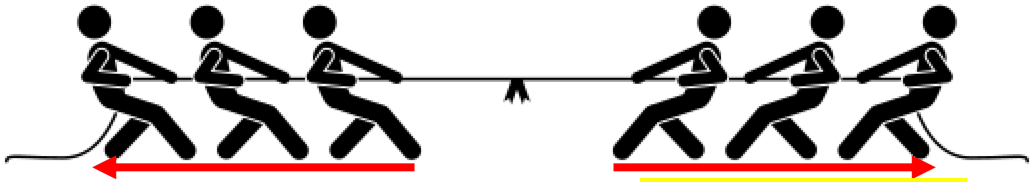


Balanced Entropy

$$-\infty \leq \mathbf{BalEnt[x]} := \frac{\sum \mathrm{E} P_i h(P_i^+) + H(Y)}{H(Y) + \log 2} \leq 1, \qquad Y = \text{Class Label}$$

$$\sum \mathrm{E} P_i h(P_i^+) \leq 0$$

✓ Continuous differential entropy

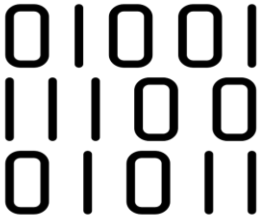


✓ Model confidence

✓ Label uncertainty

$$H(Y) \geq 0$$

✓ Discrete Shannon entropy



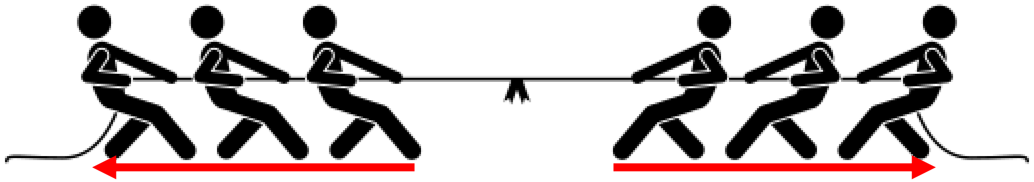
SAMSUNG SDS

Balanced Entropy

$$-\infty \leq \mathbf{BalEnt}[\mathbf{x}] := \frac{\sum \mathrm{E} P_i h(P_i^{+}) + H(Y)}{H(Y) + \log 2} \leq 1, \qquad Y = \text{Class Label}$$

$\sum \mathrm{E} P_i h(P_i^{+}) \leq 0$

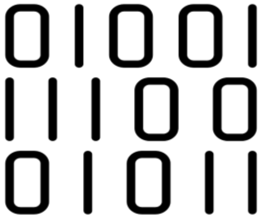
✓ Continuous differential entropy



✓ Model confidence ✓ Label uncertainty

$H(Y) \geq 0$

✓ Discrete Shannon entropy



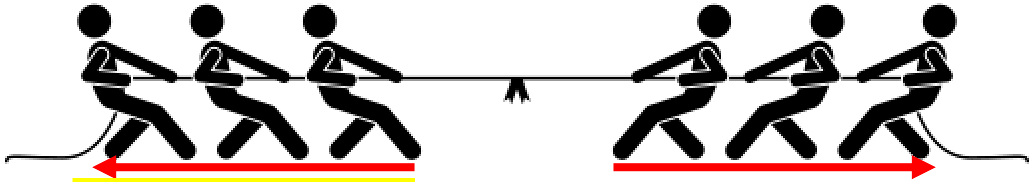
SAMSUNG SDS

Balanced Entropy

$$-\infty \leq \mathbf{BalEnt}[\mathbf{x}] := \frac{\sum \mathrm{E} P_i h(P_i^+) + H(Y)}{H(Y) + \log 2} \leq 1, \qquad Y = \text{Class Label}$$

$\sum \mathrm{E} P_i h(P_i^+) \leq 0$

✓ Continuous differential entropy

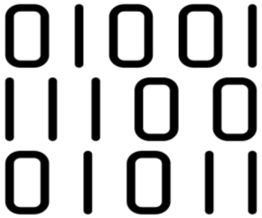


✓ Model confidence

✓ Label uncertainty

$H(Y) \geq 0$

✓ Discrete Shannon entropy



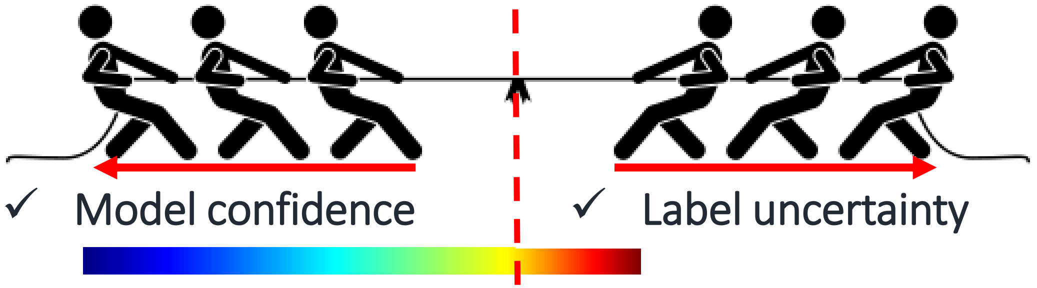
SAMSUNG SDS

Balanced Entropy

$$-\infty \leq \mathbf{BalEnt[x]} := \frac{\sum \mathrm{E} P_i h(P_i^+) + H(Y)}{H(Y) + \log 2} \leq 1, \qquad Y = \text{Class Label}$$

$\sum \mathrm{E} P_i h(P_i^+) \leq 0$

✓ Continuous differential entropy



✓ Model confidence

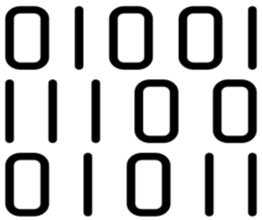
✓ Label uncertainty

$\mathbf{BalEnt[x]} \approx 0$

✓ Expected estimation error probability after acquisition

$H(Y) \geq 0$

✓ Discrete Shannon entropy



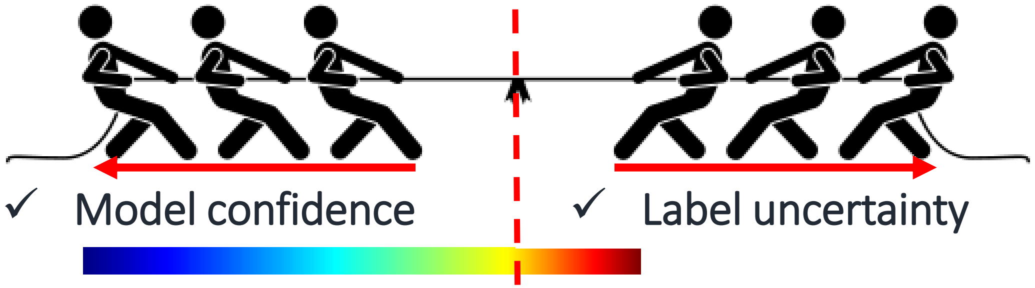
SAMSUNG SDS

Balanced Entropy

$$-\infty \leq \mathbf{BalEnt[x]} := \frac{\sum \mathrm{E} P_i h(P_i^+) + H(Y)}{H(Y) + \log 2} \leq 1, \qquad Y = \text{Class Label}$$

$\sum \mathrm{E} P_i h(P_i^+) \leq 0$

- ✓ Continuous differential entropy



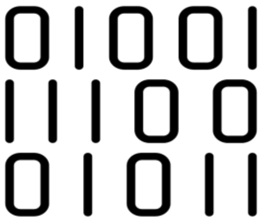
- ✓ Model confidence
- ✓ Label uncertainty

$\mathbf{BalEnt[x]} \approx 0$

- ✓ Expected estimation error probability after acquisition

$H(Y) \geq 0$

- ✓ Discrete Shannon entropy



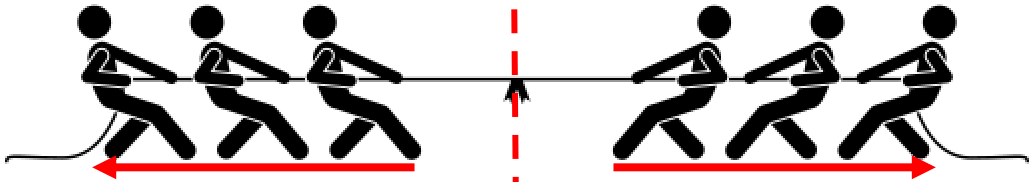
SAMSUNG SDS

Balanced Entropy

$$-\infty \leq \mathbf{BalEnt[x]} := \frac{\sum \mathrm{E} P_i h(P_i^+) + H(Y)}{H(Y) + \log 2} \leq 1, \qquad Y = \text{Class Label}$$

$\sum \mathrm{E} P_i h(P_i^+) \leq 0$

✓ Continuous differential entropy

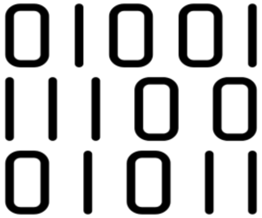


✓ Model confidence

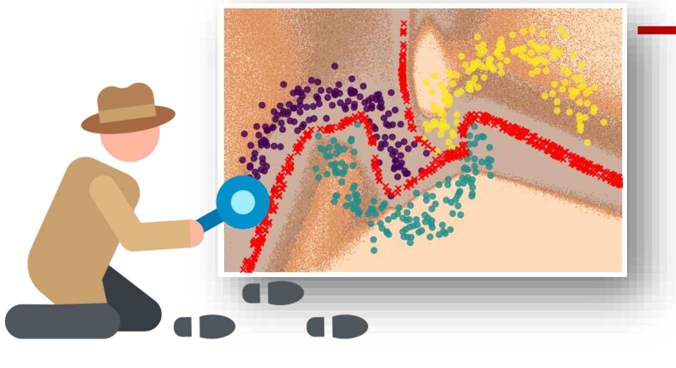
✓ Label uncertainty

$H(Y) \geq 0$

✓ Discrete Shannon entropy



$\mathbf{BalEnt[x]} \approx 0$



✓ Expected estimation error probability after acquisition

SAMSUNG SDS

Balanced Entropy

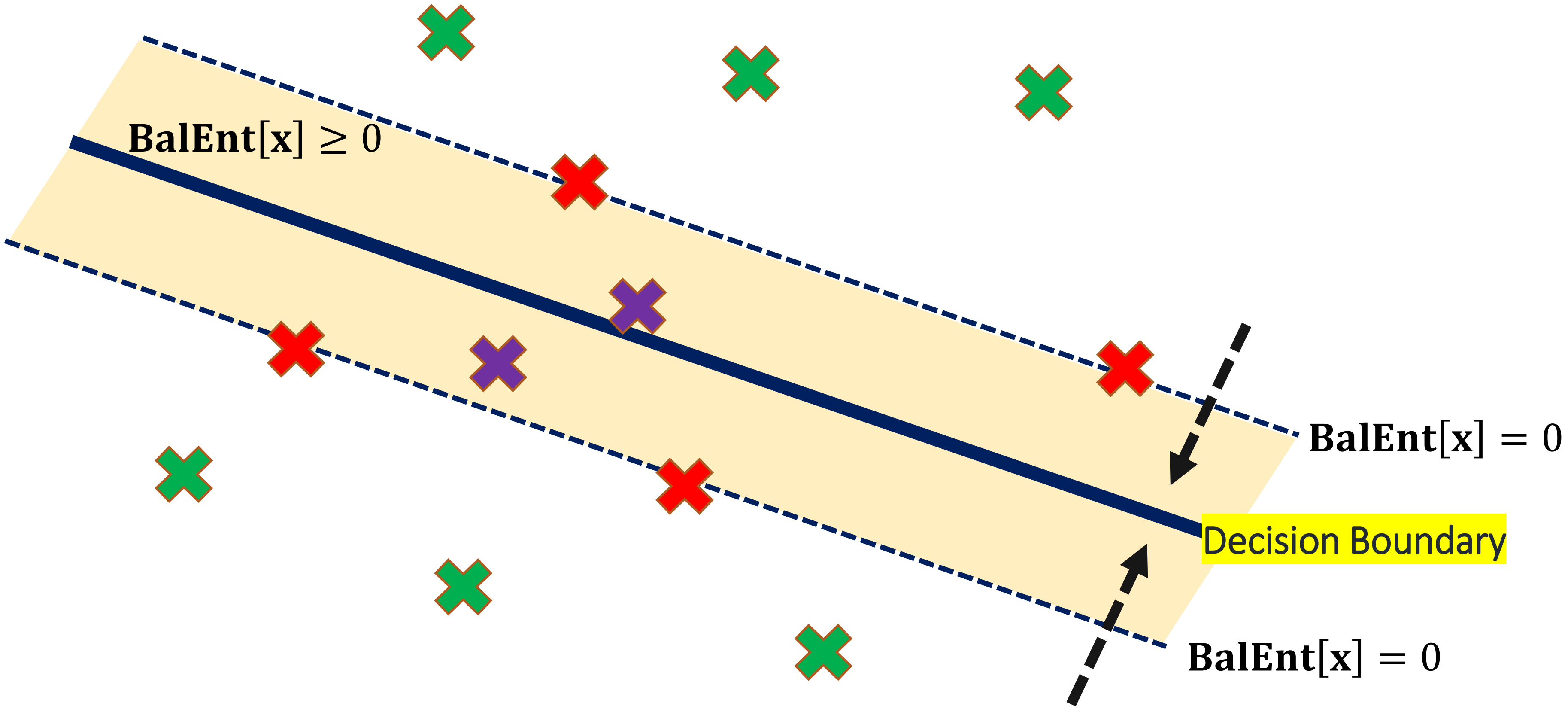
FYI,

$$\mathbf{BalEnt}[\mathbf{x}] = \frac{\sum\left(\frac{\alpha_i}{\alpha_i+\beta_i}\right)\left[\log B(\alpha_i+1,\beta_i) - \alpha_i\Psi(\alpha_i+1) - (\beta_i-1)\Psi(\beta_i) - (\alpha_i+\beta_i-1)\Psi(\alpha_i+\beta_i+1) - \log\left(\frac{\alpha_i}{\alpha_i+\beta_i}\right)\right]}{-\sum\left(\frac{\alpha_i}{\alpha_i+\beta_i}\right)\log\left(\frac{\alpha_i}{\alpha_i+\beta_i}\right) + \log 2},$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

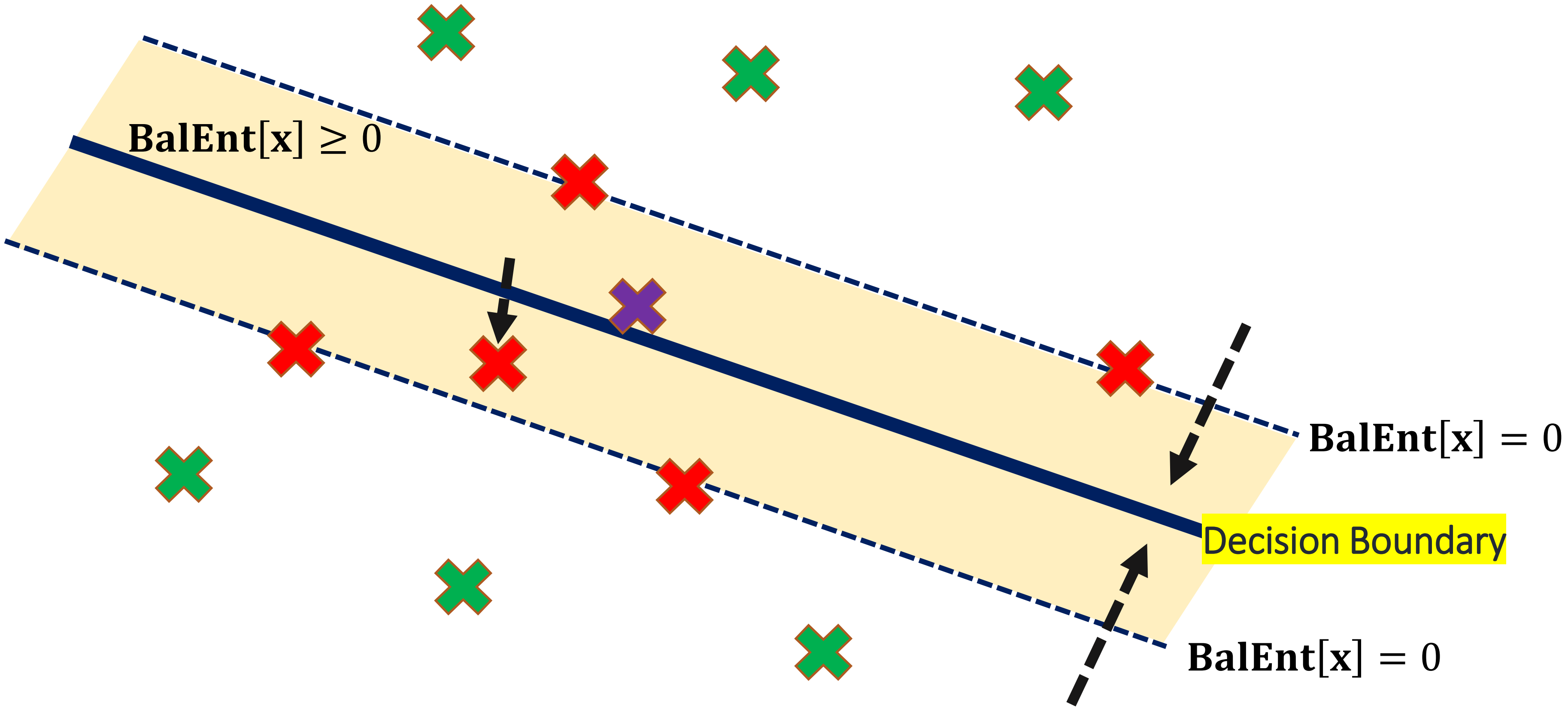
SAMSUNG SDS

Balanced Entropy Learning Principle



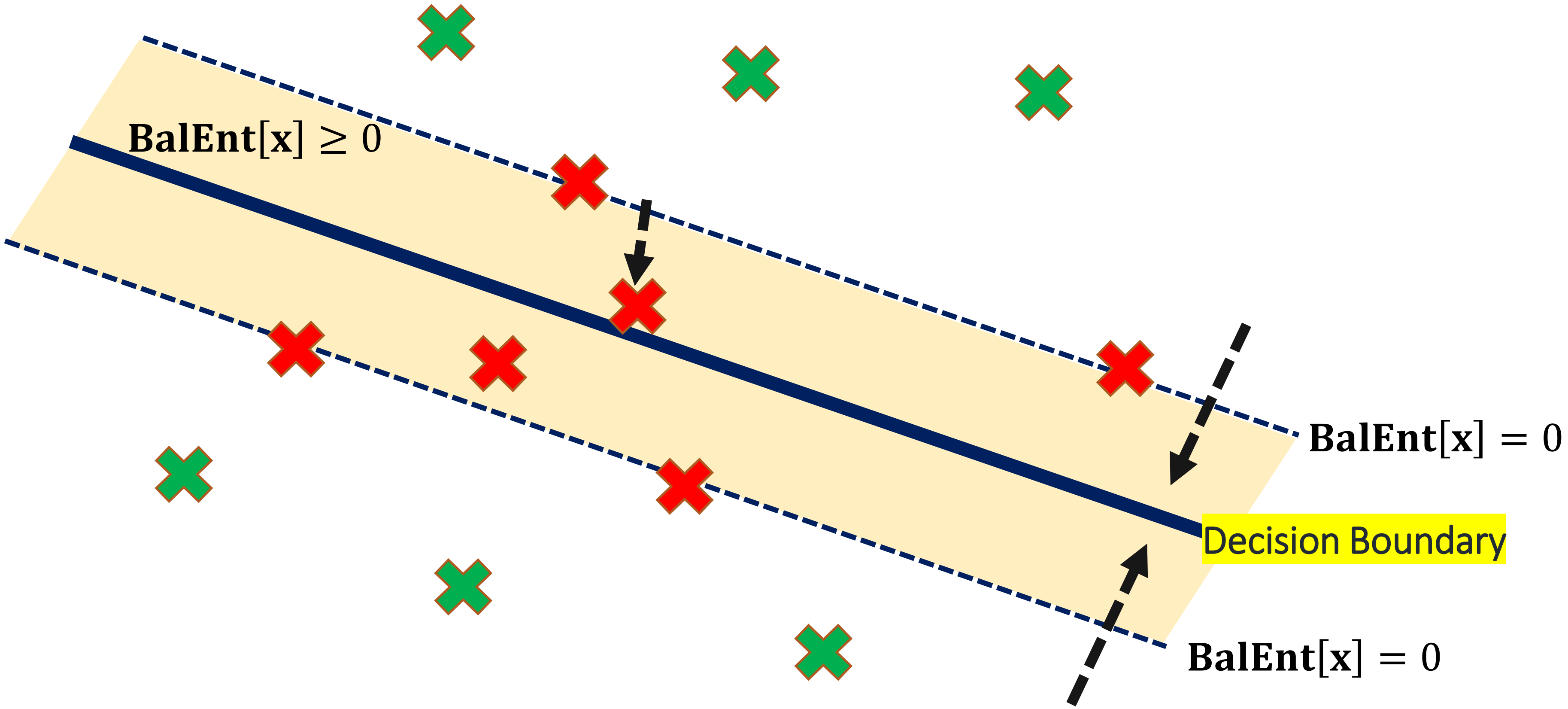
SAMSUNG SDS

Balanced Entropy Learning Principle



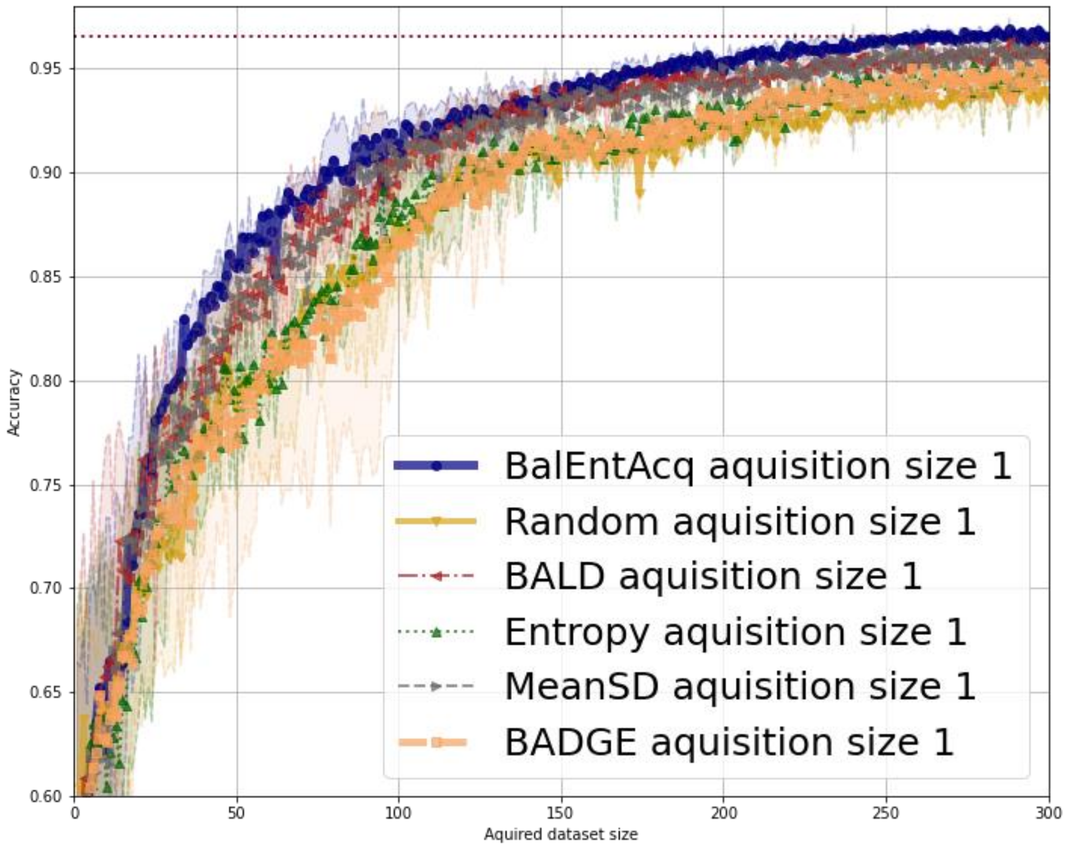
SAMSUNG SDS

Balanced Entropy Learning Principle

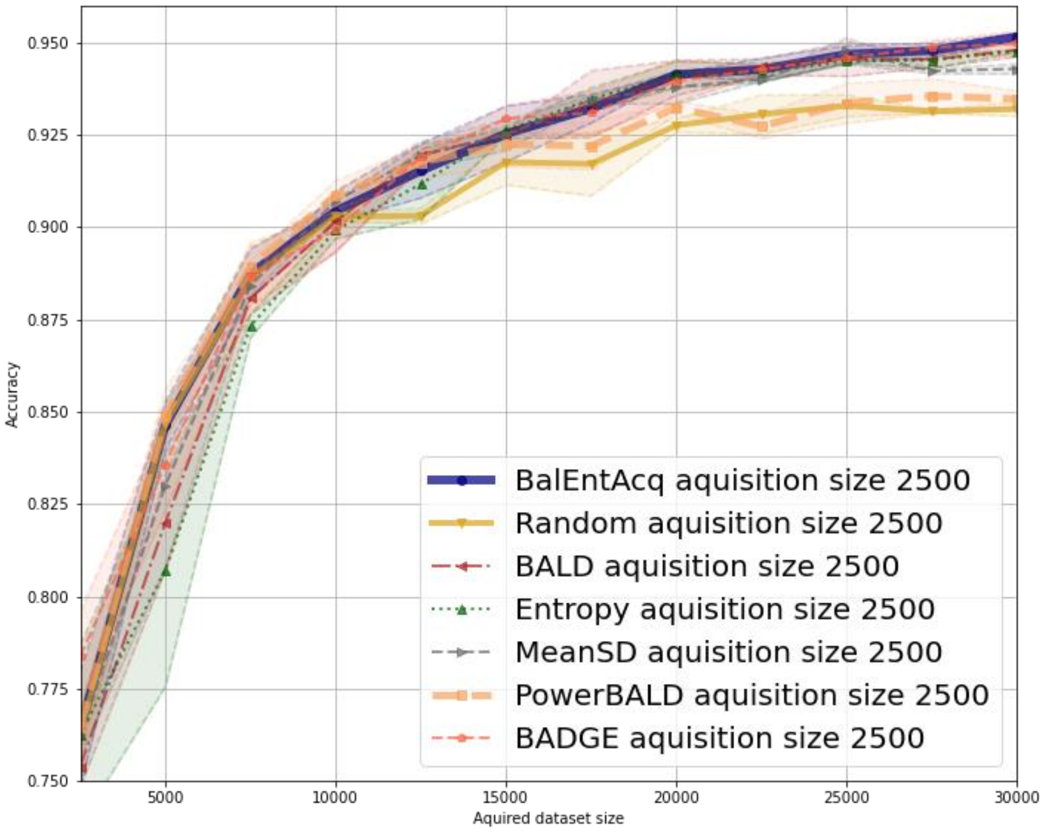


[1] Yinglun Zhu and Robert Nowak. Efficient active learning with abstention. Advances in Neural Information Processing Systems, 2022b.17
[2] Yinglun Zhu and Robert Nowak. Active learning with neural networks: Insights from nonparametric statistics. Advances in Neural Information Processing Systems, 2022a.

Results



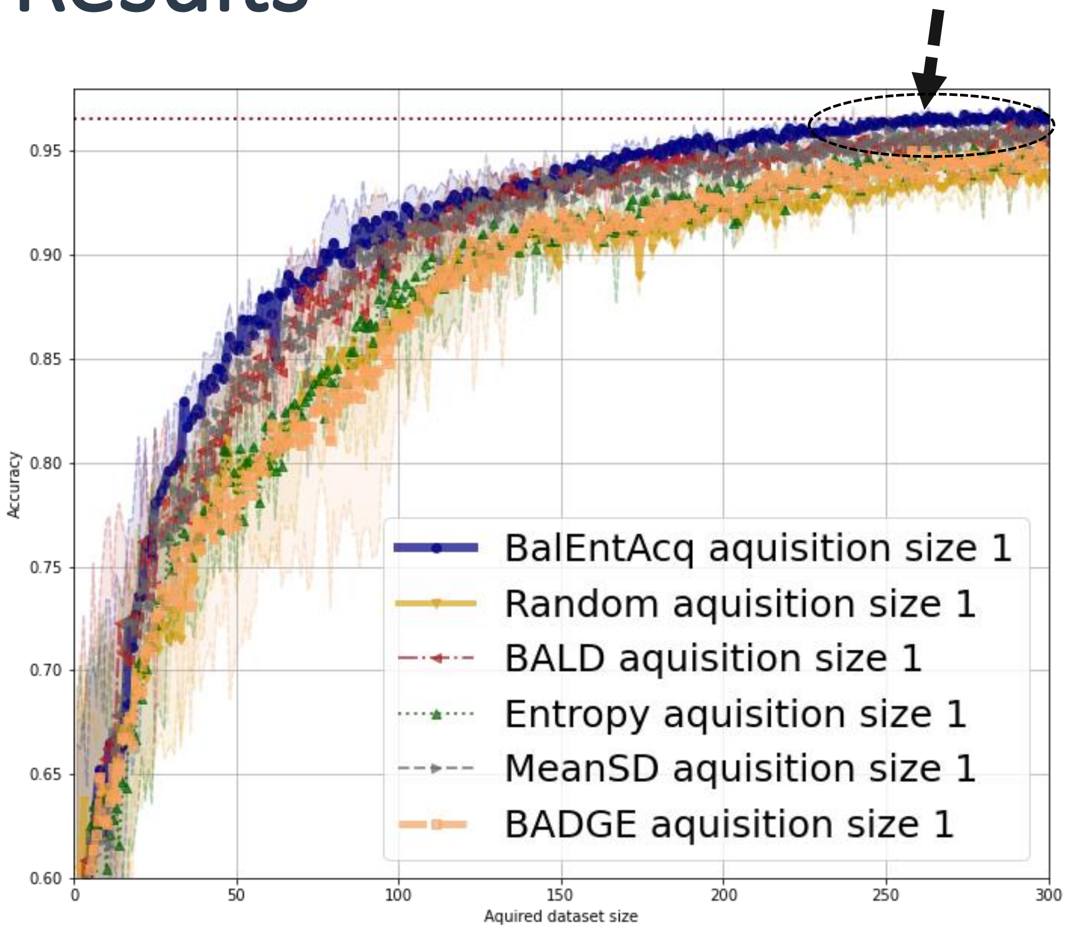
✓ MNIST



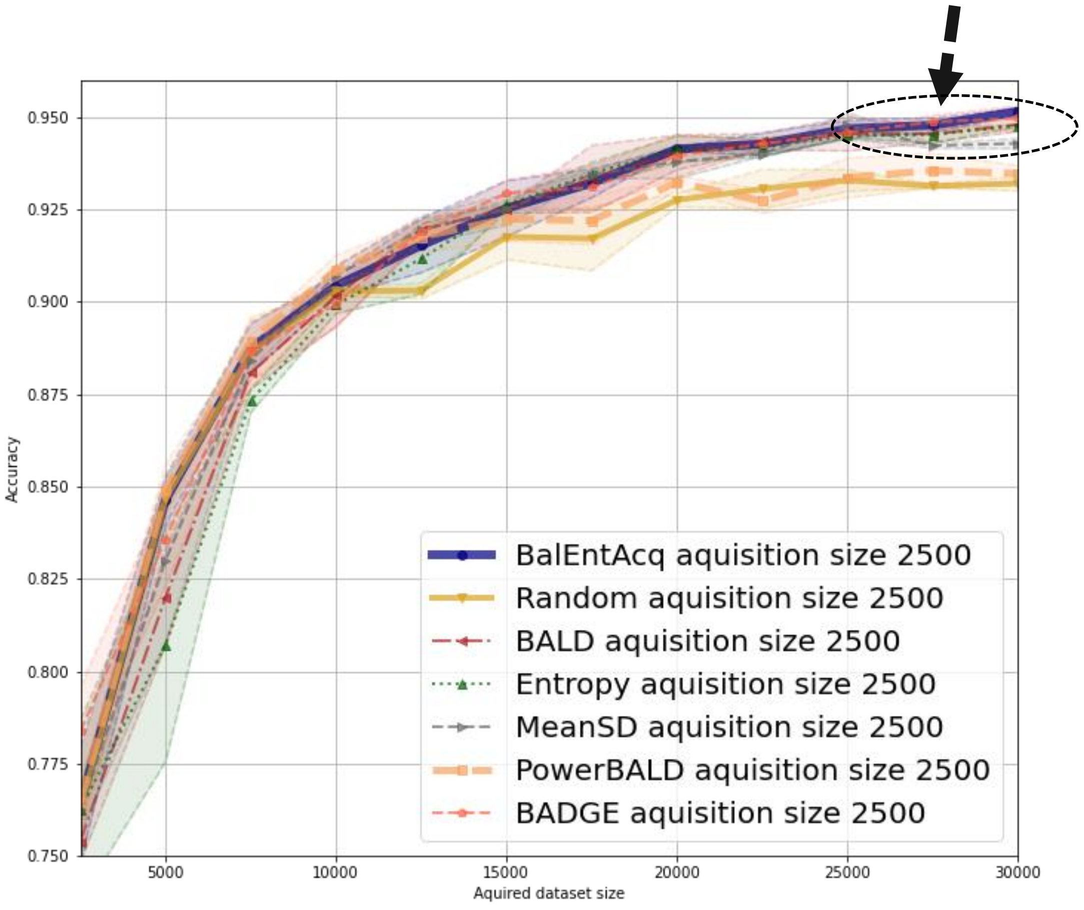
✓ SVHN

SAMSUNG SDS

Results



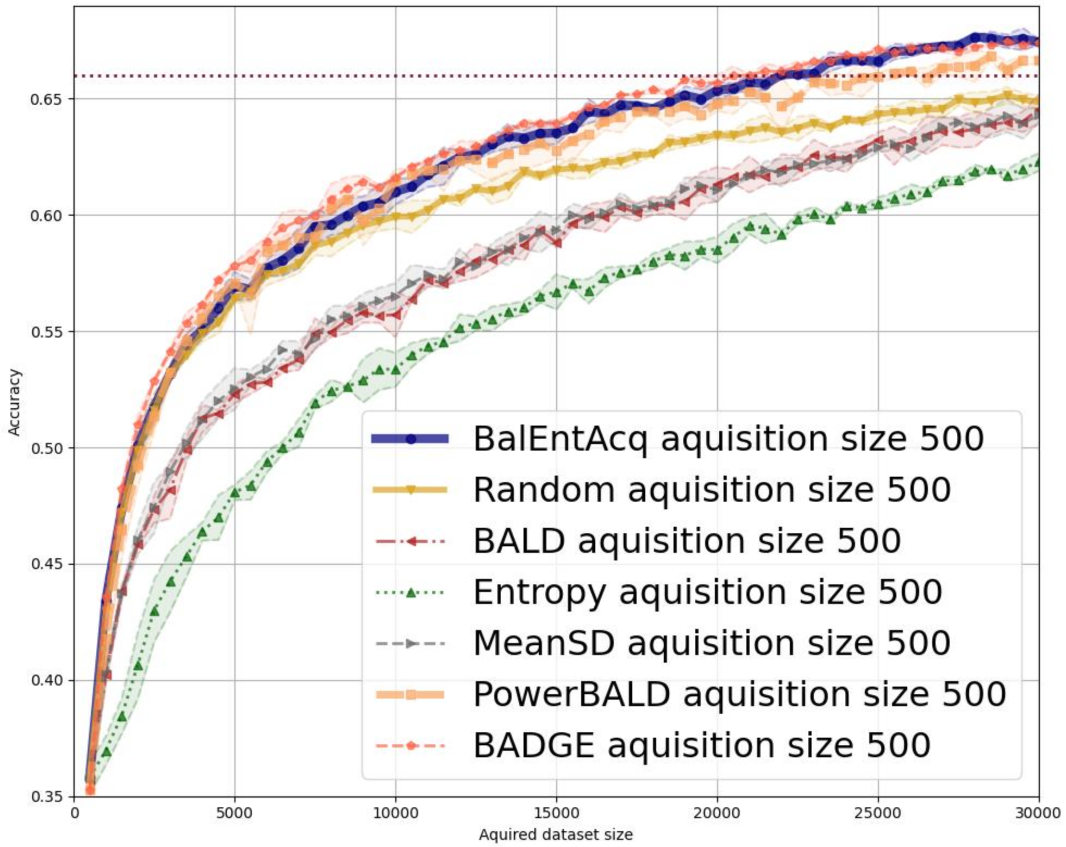
✓ MNIST



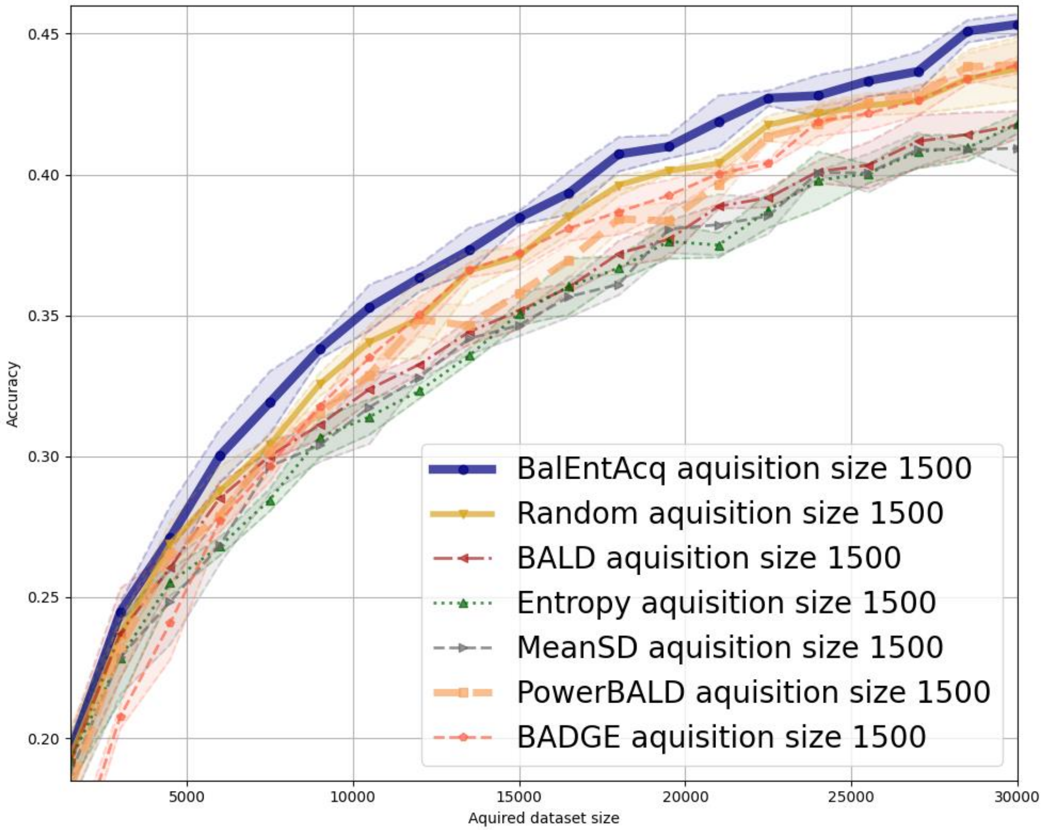
✓ SVHN

SAMSUNG SDS

Results



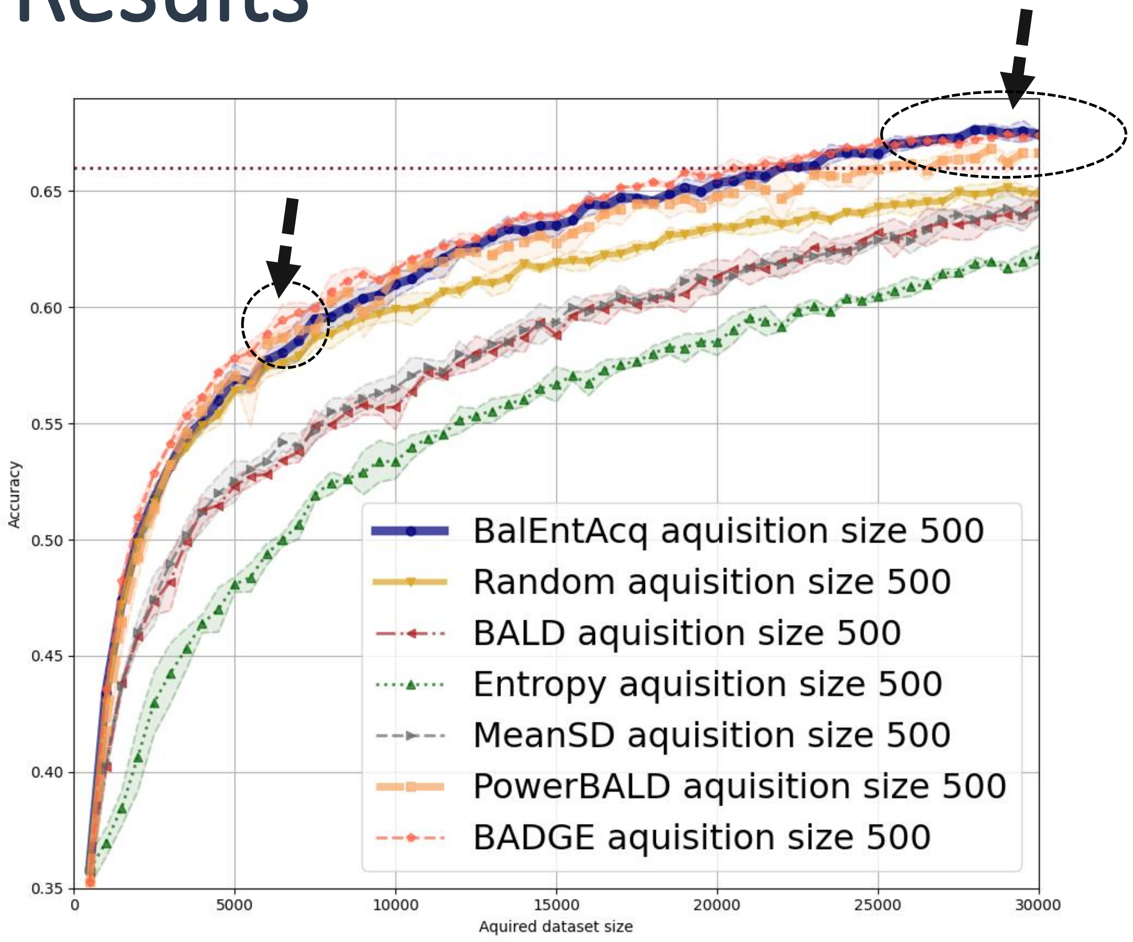
✓ 3 x CIFAR-100

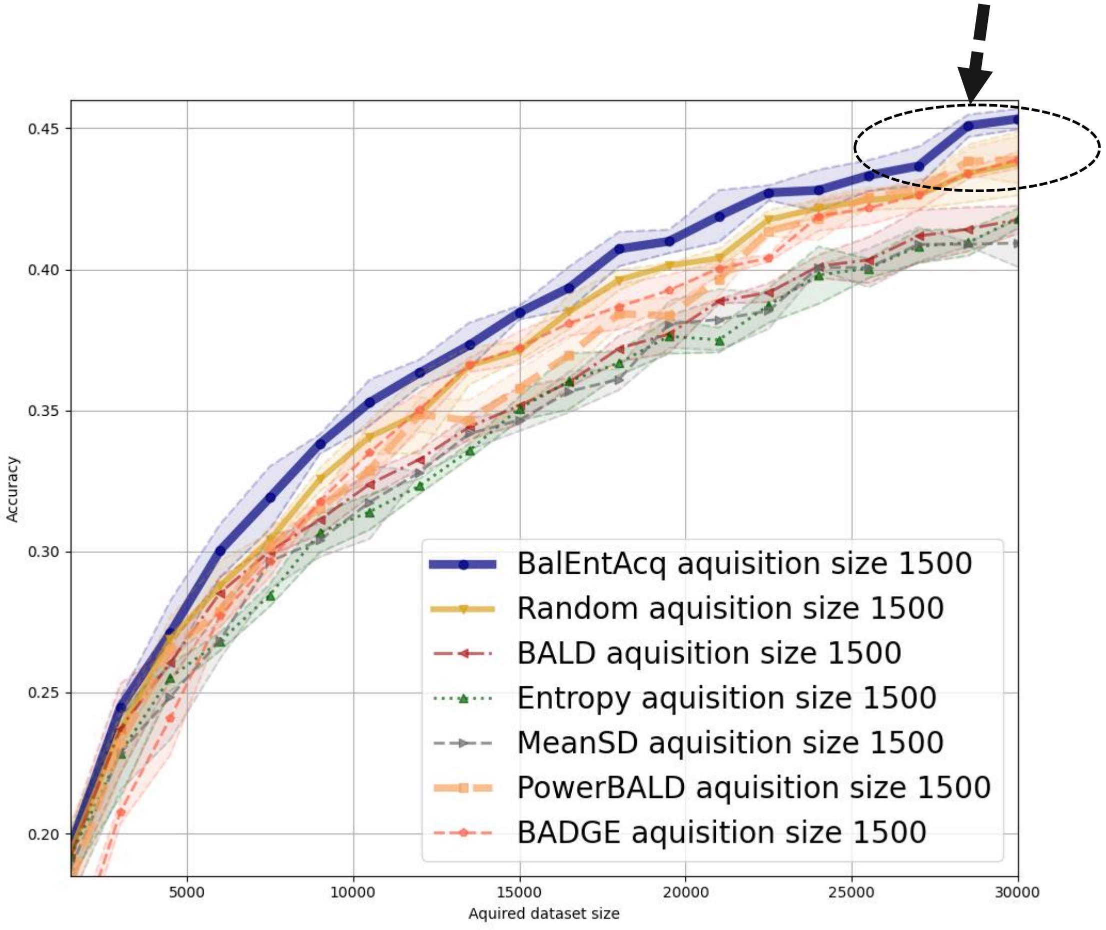


✓ Tiny-ImageNet

SAMSUNG SDS

Results





✓ 3 x CIFAR-100

✓ Tiny-ImageNet

SAMSUNG SDS

Conclusion

- ✓ Balanced Entropy provides a novel way to quantify the uncertainty in active learning
  - ✓ Not maximal information gain, but sufficient to achieve superior performance than others
  - ✓ Linear computational time complexity
  - ✓ Diversified selection
- ✓ Balanced Entropy comes from the joint entropy formulation between the model and the label
  - ✓ It quantifies the estimation error probability after acquisition under entropy precision
- ✓ Look forward to having further applications beyond active learning

SAMSUNG SDS