Learning to Benchmark

Alfred Hero

Michigan Institute for Data Science (MIDAS)
Dept of Electrical Engineering and Computer Science (EECS)
Dept of Biomedical Engineering (BME)
Dept of Statistics
University of Michigan - Ann Arbor

Dec. 20, 2018

Benchmarks and benchmarking

adobe.com

adweek.com

1. Meta-learning
2. Performance benchmarks
3. Empirical information divergence estimation
4. Ensemble divergence estimators
5. Applications
6. Summary

Predicting classification performance

The problem of predicting fundamental performance limits has long history:

- Optimal design of experiments - Gergonne[1], Smith[2], Lindley[3]

[1] J.D. Gergonne (1815), Application de la méthode des moindre quarrés a l'interpolation des suites. Annales de Mathématiques Pures et Appliquées. 5:242-252

[2] K. Smith (1918), "On the standard deviations of adjusted ... Biometrika 12:1-85.

[3] D. V. Lindley (1956). On a measure of information provided by an experiment. Annals of Math Stat., 27(4):986-1005.

Predicting performance

The problem of predicting performance limits has a long history:

- Optimal design of experiments - Gergonne[1], Smith[2], Lindley[3]
- Optimal decision processes - Belman[4], DeGroot[5], Sutton[6]

[1] J.D. Gergonne (1815), Application de la méthode des moindre quarrés a l'interpolation des suites. Annales de Mathématiques Pures et Appliquées. 5:242-252

[2] K. Smith (1918), "On the standard deviations of adjusted ... Biometrika 12:1-85.

[3] D. V. Lindley (1956). On a measure of information provided by an experiment. Annals of Math Stat., 27(4):986-1005.

[4] R. Bellman (1957). A Markovian Decision Process. Journal of Mathematics and Mechanics.

[5] M. DeGroot Optimal Statistical Decisions. McGraw-Hill. New York. 1970.

[6] R.S. Sutton and A.G. Barto(1998). Reinforcement Learning: An Introduction. MIT Press.

Predicting performance

The problem of predicting performance limits has a long history:

- Optimal design of experiments - Gergonne[1], Smith[2], Lindley[3]
- Optimal decision processes - Belman[4], DeGroot[5], Sutton[6]
- Model selection - Akaike[7], Stein[8], Donoho[9].

---

[1] J.D. Gergonne (1815), Application de la méthode des moindre quarrés a l'interpolation des suites. Annales de Mathématiques Pures et Appliquées. 5:242-252

[2] K. Smith (1918), "On the standard deviations of adjusted ... Biometrika 12:1-85.

[3] D. V. Lindley (1956). On a measure of information provided by an experiment. Annals of Math Stat., 27(4):986-1005.

[4] R. Bellman (1957). A Markovian Decision Process. Journal of Mathematics and Mechanics.

[5] M. DeGroot Optimal Statistical Decisions. McGraw-Hill. New York. 1970.

[6] R.S. Sutton and A.G. Barto(1998). Reinforcement Learning: An Introduction. MIT Press.

[7] Akaike, H. (1974). A new look at the statistical model identification. IEEE T. Autom. control, 19(6):716-723.

[8] C.M Stein (1981), Estimation of the mean of a multivariate normal distribution. Annals of Statistics, 9(6):1135-1151.

[9] D.L. Donoho and I.M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. J. Amer Stat. Assoc, 90(432):1200-1224.

Learning performance limits by empirical estimation of Bayes error bound

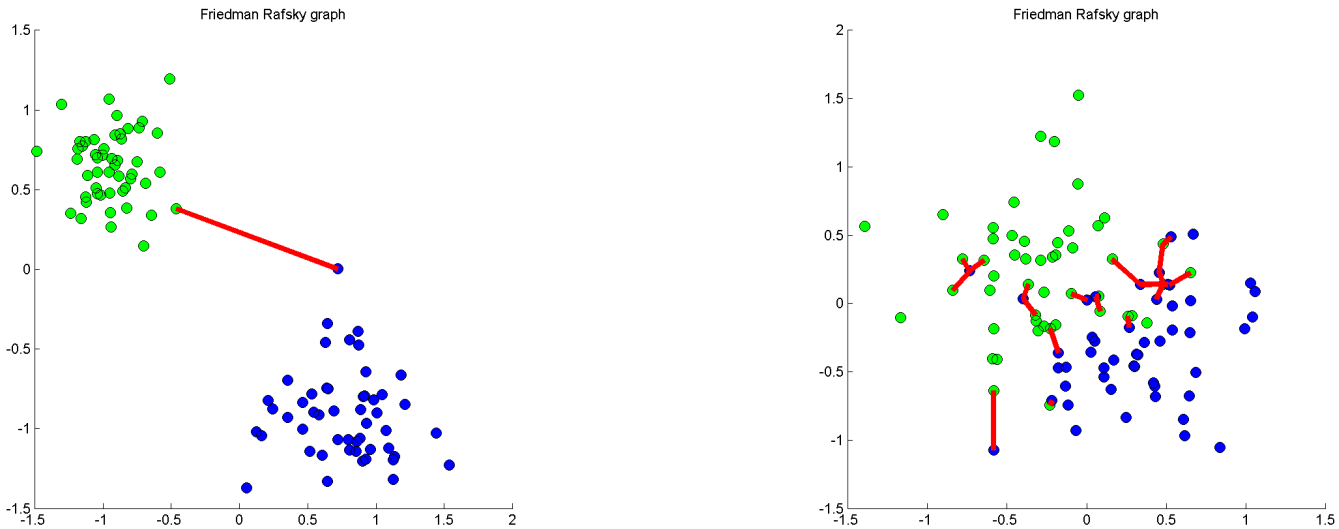


Figure: Friedman-Rafsky statistic converge to bounds on Bayes classification error.

- Q: Can we empirically estimate tight upper and lower bounds on Bayes misclassification error?

[1] J. Friedman and L. Rafsky (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. The Annals of Statistics: 697-717.

[2] N. Henze and M. D. Penrose (1999). On the multivariate runs test. Annals of Stats, 290-298.

[3] M. Noshad and A. Hero (2018). Rate-optimal meta learning of classification error. IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc. (ICASSP)

Learning performance limits by empirical estimation of Bayes error bound

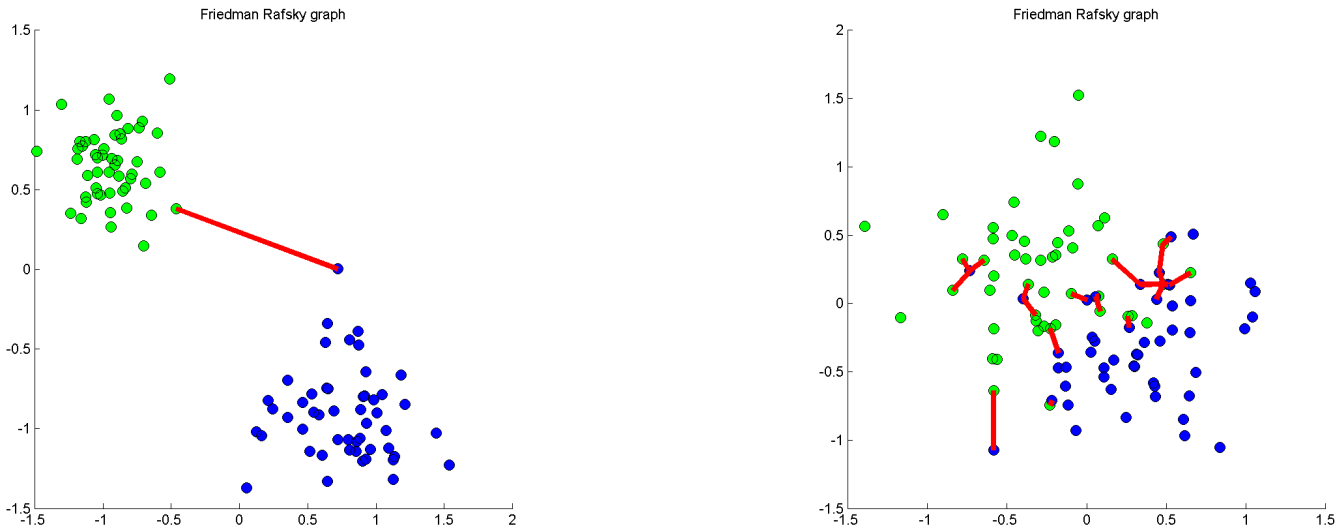


Figure: Friedman-Rafsky statistic converge to bounds on Bayes classification error.

- Q: Can we empirically estimate tight upper and lower bounds on Bayes misclassification error?
- A: Boosted ensembles of modified FR information divergence estimators[1] are rate-optimal estimates of a divergence measure[2] for smooth distributions[3]

[1] J. Friedman and L. Rafsky (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. The Annals of Statistics: 697-717.

[2] N. Henze and M. D. Penrose (1999). On the multivariate runs test. Annals of Stats, 290-298.

[3] M. Noshad and A. Hero (2018). Rate-optimal meta learning of classification error. IEEE Intl. Conf. on Acoust., Speech, and Sig. Proc. (ICASSP)

Bayes optimal classifier

Consider classification problem

- $Y \in \{1, \ldots, K\}$ an unknown label with priors $\{p_k\}$, $\sum_{k=1}^{K} p_k = 1$.

$$P(Y = k) = \begin{cases} p_1, & k = 1 \\ \vdots, & \vdots \\ p_K, & k = K \end{cases}$$

- X an observed random variable with conditional distribution

$$f(x|Y = k) = \begin{cases} f_1(x), & k = 1 \\ \vdots, & \vdots \\ f_K(x), & k = K \end{cases}$$

Bayes optimal classifier

Consider classification problem

- $Y \in \{1, \ldots, K\}$ an unknown label with priors $\{p_k\}$, $\sum_{k=1}^{K} p_k = 1$.

$$P(Y = k) = \begin{cases} p_1, & k = 1 \\ \vdots, & \vdots \\ p_K, & k = K \end{cases}$$

- X an observed random variable with conditional distribution

$$f(x|Y = k) = \begin{cases} f_1(x), & k = 1 \\ \vdots, & \vdots \\ f_K(x), & k = K \end{cases}$$

Let $C(x)$ be (Bayes) optimal classifier that minimizes avg 0-1 loss (probability of error)

$$C(x) = \text{argmax}_{1 \leq k \leq K} P(Y = k|X = x)$$

Bayes error rate: best achievable misclassification error probability

Bayes error rate is avg missclassification error probability of Bayes classifier

$$\epsilon_{p_1,\ldots,p_K}(f_1, f_2, \ldots, f_K) = P(C(X) \neq Y)$$

Bayes error rate: best achievable misclassification error probability

Bayes error rate is avg missclassification error probability of Bayes classifier

$$\epsilon_{p_1,\ldots,p_K}(f_1, f_2, \ldots, f_K) = P(C(X) \neq Y)$$

Integral representation

$$\epsilon_{p_1,\ldots,p_K}(f_1, f_2, \ldots, f_K) = 1 - \int \max_{1 \leq k \leq K} \{f_k(x) p_k\} dx$$

Bayes error rate: best achievable misclassification error probability

Bayes error rate is avg missclassification error probability of Bayes classifier

$$\epsilon_{p_1,\ldots,p_K}(f_1, f_2, \ldots, f_K) = P(C(X) \neq Y)$$

Integral representation

$$\epsilon_{p_1,\ldots,p_K}(f_1, f_2, \ldots, f_K) = 1 - \int \max_{1 \leq k \leq K} \{f_k(x) p_k\} dx$$

Benchmark learning objective: empirically estimate the Bayes error rate

- directly from n training samples $\{(X_i, Y_i)\}_{i=1}^n$
- assuming (X_i, Y_i) i.i.d., where
- given Y_i: $X_i \sim f(x|Y_i = k) = f_k(x)$, $k = 1, \ldots, K$
- $P(Y_i = k) = p_k$

Bayes error rate for binary case ($Y \in \{0, 1\}$)

For special case of binary classification ($K = 2$)

$$\epsilon_{p,q}(f_1, f_2) = \frac{1}{2} - \frac{1}{2} \int |pf_1(x) - qf_2(x)| dx,$$

where $q = 1 - p$

Bayes error rate for binary case ($Y \in \{0, 1\}$)

For special case of binary classification ($K = 2$)

$$\epsilon_{p,q}(f_1, f_2) = \frac{1}{2} - \frac{1}{2} \int |pf_1(x) - qf_2(x)| dx,$$

where $q = 1 - p$

Alternative representation:

$$\epsilon_{p_1,p_2}(f_1, f_2) = \frac{1 + |p - q|}{2} - \frac{1}{2} \int g(f_1(x)/f_2(x)) f_2(x) dx,$$

where $g(u)$ is the convex function

$$g(u) = |pu - q| - |p - q|$$

The f-divergence between a pair of distributions

The f-divergence (Csiszár)[1], (Ali-Silvey)[2]:

$$D_g(f_1 \| f_2) = \int g\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x) dx$$

where $g(u)$ is a convex function on $\mathbb{R}^+$ and $g(1) = 0$.

[1] I. Csiszár (1963), Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. Magyar. Tud. Akad. Mat. Kutato Int. Kozl. 8:85108.

[2] S. M. Ali and S. D. Silvey (1966), A general class of coefficients of divergence of one distribution from another, J. Royal Stat. Soc., Ser.B , 28:131-142.

The f-divergence between a pair of distributions

The f-divergence (Csiszár)[1], (Ali-Silvey)[2]:

$$D_g(f_1 \| f_2) = \int g\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x) dx$$

where $g(u)$ is a convex function on $\mathbb{R}^+$ and $g(1) = 0$.

Properties: if g is strictly convex then $D_g(f_1 \| f_2)$ is

- reflexive non-negative: $D_g(f_1 \| f_2) \geq 0$ with equality iff $f_1 = f_2$
- monotone: $D_g(f_1 \| f_2)$ non-increasing under transformations $x \to T(x)$
- jointly convex: $D_g(f_1 \| f_2)$ is convex in (f_1, f_2)

[1] I. Csiszár (1963), Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. Magyar. Tud. Akad. Mat. Kutato Int. Kozl. 8:85108.

[2] S. M. Ali and S. D. Silvey (1966), A general class of coefficients of divergence of one distribution from another, J. Royal Stat. Soc., Ser.B , 28:131-142.

Instances of f-divergences

The following are common instances of f-divergence[1]

- Total variation distance $g(u) = \frac{1}{2}|u-1|$

$$D^{TV}(f_1 \| f_2) = \frac{1}{2}\int |f_1(x) - f_2(x)| dx$$

[1] Csiszár, I., and Shields, P. C. (2004). Information theory and statistics: A tutorial. Foundations and Trends in Communications and Information Theory, 1(4), 417-528.

Instances of f-divergences

The following are common instances of f-divergence[1]

- Total variation distance $g(u) = \frac{1}{2}|u-1|$

$$D^{TV}(f_1 \| f_2) = \frac{1}{2}\int |f_1(x) - f_2(x)| dx$$

- α-divergence: $g(u) = (1-u^{\alpha})\frac{1}{1-\alpha}$

$$D^{R}(f_1 \| f_2) = \left(1 - \int f_1^{\alpha}(x) f_2^{1-\alpha}(x) dx\right) \frac{1}{1-\alpha}$$

[1] Csiszár, I., and Shields, P. C. (2004). Information theory and statistics: A tutorial. Foundations and Trends in Communications and Information Theory, 1(4), 417-528.

Instances of f-divergences

The following are common instances of f-divergence[1]

- Total variation distance $g(u) = \frac{1}{2}|u-1|$

$$D^{TV}(f_1 \| f_2) = \frac{1}{2}\int |f_1(x) - f_2(x)| dx$$

- α-divergence: $g(u) = (1-u^{\alpha})\frac{1}{1-\alpha}$

$$D^{R}(f_1 \| f_2) = \left(1 - \int f_1^{\alpha}(x) f_2^{1-\alpha}(x) dx\right) \frac{1}{1-\alpha}$$

- Kullback-Liebler divergence: $g(u) = u\log u$:

$$D^{KL}(f_1 \| f_2) = \int f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right) dx$$

[1] Csiszár, I., and Shields, P. C. (2004). Information theory and statistics: A tutorial. Foundations and Trends in Communications and Information Theory, 1(4), 417-528.

Instances of f-divergences

The following are common instances of f-divergence[1]

- Total variation distance $g(u) = \frac{1}{2}|u-1|$

$$D^{TV}(f_1\|f_2) = \frac{1}{2}\int |f_1(x) - f_2(x)|dx$$

- α-divergence: $g(u) = (1-u^{\alpha})\frac{1}{1-\alpha}$

$$D^{R}(f_1\|f_2) = \left(1 - \int f_1^{\alpha}(x) f_2^{1-\alpha}(x)dx\right)\frac{1}{1-\alpha}$$

- Kullback-Liebler divergence: $g(u) = u\log u$:

$$D^{KL}(f_1\|f_2) = \int f_1(x)\log\left(\frac{f_1(x)}{f_2(x)}\right)dx$$

- Hellinger-Bhattacharyya divergence $g(u) = (\sqrt{u}-1)^2$

$$D^{H}(f_1\|f_2) = \int \left(\sqrt{f_1(x)} - \sqrt{f_2(x)}\right)^2 dx$$

[1] Csiszár, I., and Shields, P. C. (2004). Information theory and statistics: A tutorial. Foundations and Trends in Communications and Information Theory, 1(4), 417-528.

Other instances of f-divergences

- Generalized total variation distance[1]: $g(u) = |pu - q|/2 - |p - q|/2$

$$D_p^{GTV} = \frac{1}{2} \int |pf_1(x) - qf_2(x)| dx + |p - q|/2$$

- Henze-Penrose divergence[2]: $g(u) = \frac{1}{4pq} \left[\frac{(pt-q)^2}{pt+q} - (p-q)^2 \right]$

$$D_p^{HP} = \frac{1}{4pq} \left[\int \frac{(pf_1(x) - qf_2(x))^2}{pf_1(x) + qf_2(x)} dx - (p-q)^2 \right].$$

[1] T. Kailath (1967), The divergence and Bhattacharyya distance measures in signal selection, IEEE T. Communication Technology, 15:1:5260

[2] N. Henze and M. D. Penrose (1999). On the multivariate runs test. Annals of Stats, 290-298.

f-divergences and Bayes error rate

These divergences can each be related to minimum probability of error

- Exact f-divergence representation

$$\epsilon_{p,q}(f_1, f_2) = \frac{1 + |p - q|}{2} - D_p^{GTV}(f_1(x) \| f_2(x))$$

[1] T. Kailath (1967), The divergence and Bhattacharyya distance measures in signal selection, IEEE T. Communication Technology, 15:1:5260

[2] Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. arXiv preprint physics/0004057.

[3] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning (pp. 1096-1103). ACM.

f-divergences and Bayes error rate

These divergences can each be related to minimum probability of error

- Exact f-divergence representation

$$\epsilon_{p,q}(f_1, f_2) = \frac{1 + |p - q|}{2} - D_p^{GTV}(f_1(x) \| f_2(x))$$

- Bhattacharyya bound[1]

$$\frac{1}{2} - \frac{1}{2}\sqrt{1 - BC_p^2} \leq \epsilon_{p,q} \leq \frac{1}{2} BC_p,$$

where $BC_p = \frac{\sqrt{pq}}{2}(1 - D_p^H)$ is the Bhattacharyya coefficient

[1] T. Kailath (1967), The divergence and Bhattacharyya distance measures in signal selection, IEEE T. Communication Technology, 15:1:5260

[2] Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. arXiv preprint physics/0004057.

[3] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning (pp. 1096-1103). ACM.

f-divergences and Bayes error rate

These divergences can each be related to minimum probability of error

- Exact f-divergence representation

$$\epsilon_{p,q}(f_1, f_2) = \frac{1 + |p - q|}{2} - D_p^{GTV}(f_1(x) \| f_2(x))$$

- Bhattacharyya bound[1]

$$\frac{1}{2} - \frac{1}{2}\sqrt{1 - BC_p^2} \leq \epsilon_{p,q} \leq \frac{1}{2} BC_p,$$

where $BC_p = \frac{\sqrt{pq}}{2}(1 - D_p^H)$ is the Bhattacharyya coefficient

- *Learning to benchmark* can be reduced to f-divergence estimation.
- f-divergences are widely used in signal processing[2] and machine learning[3].
- Density plug in estimation strategies are often applied.

[1] T. Kailath (1967), The divergence and Bhattacharyya distance measures in signal selection, IEEE T. Communication Technology, 15:1:5260

[2] Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. arXiv preprint physics/0004057.

[3] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning (pp. 1096-1103). ACM.

Henze-Penrose divergence bound on Bayes error rate

Bound based on the Henze-Penrose divergence[1]

$$\frac{1}{2} - \frac{1}{2}\sqrt{\widetilde{D}_p^{HP}(f_1, f_2)} \leq \epsilon_p \leq \frac{1}{2} - \frac{1}{2}\widetilde{D}_p^{HP}(f_1, f_2), \tag{1}$$

where $\widetilde{D}_p^{HP} = 4pqD_p^{HP} + (p-q)^2$.

[1] V. Berisha, A. Wisler, A.O. Hero, and A. Spanias (2016), Empirically Estimable Classification Bounds Based on a Nonparametric Divergence Measure IEEE T. Signal Processing, 64:3:580-591.

Henze-Penrose divergence bound on Bayes error rate

Bound based on the Henze-Penrose divergence[1]

$$\frac{1}{2} - \frac{1}{2}\sqrt{\widetilde{D}_p^{HP}(f_1, f_2)} \leq \epsilon_p \leq \frac{1}{2} - \frac{1}{2}\widetilde{D}_p^{HP}(f_1, f_2), \tag{1}$$

where $\widetilde{D}_p^{HP} = 4pqD_p^{HP} + (p-q)^2$.

- For $p = 1/2$ this is a tighter bound than the Bhattacharyya bound.

[1] V. Berisha, A. Wisler, A.O. Hero, and A. Spanias (2016), Empirically Estimable Classification Bounds Based on a Nonparametric Divergence Measure IEEE T. Signal Processing, 64:3:580-591.

Numerical comparisons[1]

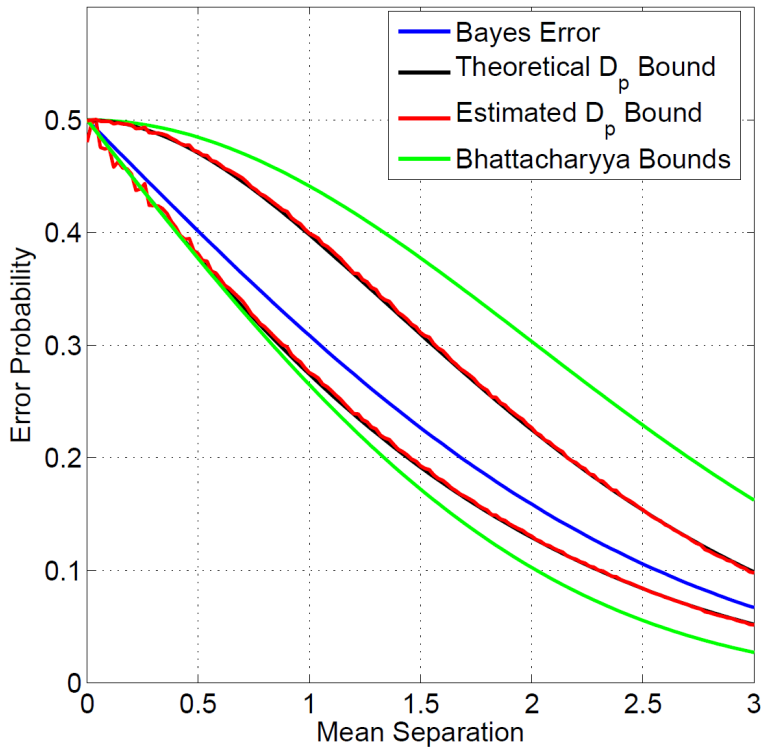


Figure: The HP bound using D_p is tighter than the Bhattacharrya bound using BC for bivariate normal distribution.

[1] V. Berisha, A. Wisler, A.O. Hero, and A. Spanias (2016), Empirically Estimable Classification Bounds Based on a Nonparametric Divergence Measure IEEE T. Signal Processing, 64:3:580-591.

Information divergence estimation

Set up:

- Given: samples $\{(X_i, Y_i)\}_{i=1}^{n}$ from $f_k(x)P(Y = k)$

$\Rightarrow$ $f_k(x) = f(x|Y = k)$

Information divergence estimation

Set up:

- Given: samples $\{(X_i, Y_i)\}_{i=1}^n$ from $f_k(x)P(Y = k)$

$\Rightarrow$ $f_k(x) = f(x|Y = k)$

- Objective: estimate divergence measure $D = D(f_1 \| f_2)$ with high accuracy

Information divergence estimation

Set up:

- Given: samples $\{(X_i, Y_i)\}_{i=1}^n$ from $f_k(x)P(Y = k)$

$\Rightarrow$ $f_k(x) = f(x|Y = k)$

- Objective: estimate divergence measure $D = D(f_1 \| f_2)$ with high accuracy

Two main approaches to estimation of D

- Density-plug-in: construct density estimators $\hat{f}_1$, $\hat{f}_2$ and form $\hat{D} = D(\hat{f}_1, \hat{f}_2)$
- Direct methods: find a simpler statistic $\hat{D}$ that converges to $D(f_1 \| f_2)$

Information divergence estimation

Set up:

- Given: samples $\{(X_i, Y_i)\}_{i=1}^n$ from $f_k(x)P(Y = k)$

$\Rightarrow$ $f_k(x) = f(x|Y = k)$

- Objective: estimate divergence measure $D = D(f_1 \| f_2)$ with high accuracy

Two main approaches to estimation of D

- Density-plug-in: construct density estimators $\hat{f}_1$, $\hat{f}_2$ and form $\hat{D} = D(\hat{f}_1, \hat{f}_2)$
- Direct methods: find a simpler statistic $\hat{D}$ that converges to $D(f_1 \| f_2)$

Estimation strategies compared along two dimensions

- Sample complexity: rate γ at which estimator RMSE decreases in n

$$RMSE = \sqrt{\mathbb{E}[(\hat{D} - D)^2]} = O(1/n^\gamma)$$

- Computational complexity: measured by memory and runtime requirements

Sample complexity for divergence estimators

Assume that $X \in \mathbb{R}^d$, $0 < c_l \leq f_X \leq c_u < \infty$, and f_X in a smooth class $\mathcal{H}_{s,K}$

Convergence results for divergence estimation:

- Minimax estimators of density functionals [1],[2]
- Minimax estimators of divergence functionals [3]
- Common behavior of minimax RMSE

$$\inf_{\widehat{D}} \sup_{f_X \in \mathcal{H}_{s,K}} E\left[\left|D(X) - \widehat{D}(X_{1:n})\right|^2\right]^{1/2} = \begin{cases} cn^{-\phi(s,d)} & s < d/m \\ cn^{-1/2} & s \geq d/2 \end{cases}$$

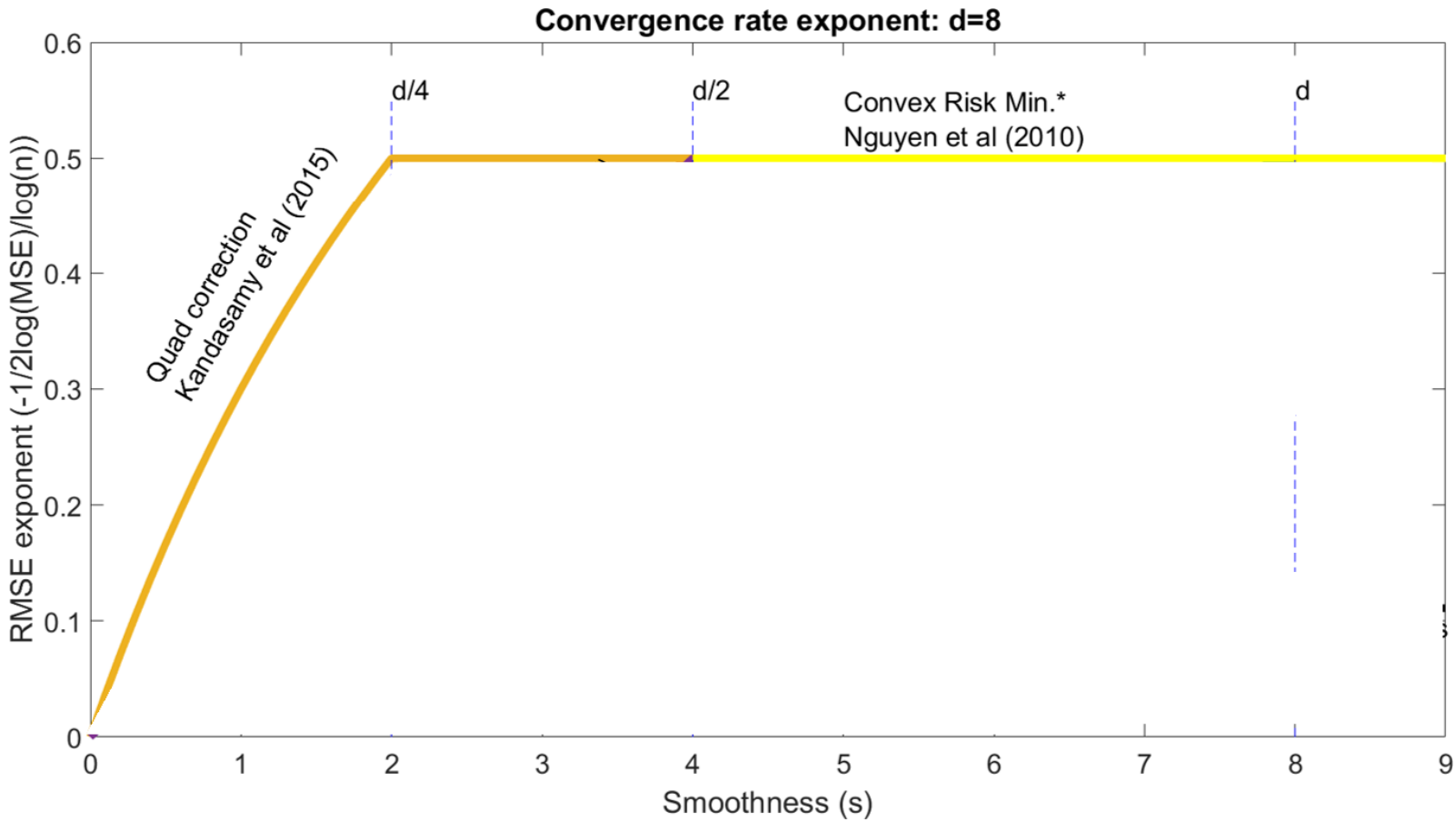
where $\phi(s,d)$ increases in s and decreases in d, and m is integer

[1] Birgé, L., and Massart, P. (1995). Estimation of integral functionals of a density. The Annals of Statistics, 11-29.

[2] Bickel, P. J., and Ritov, Y. A. (2003). Nonparametric estimators which can be "plugged-in". Annals of Statistics, 1033-1053.

[3] Nguyen, X., Wainwright, M., and Jordan, M. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE T. on Inform Theory, 56(11), 5847-5861.

RMSE convergence rate comparisons[1][2]

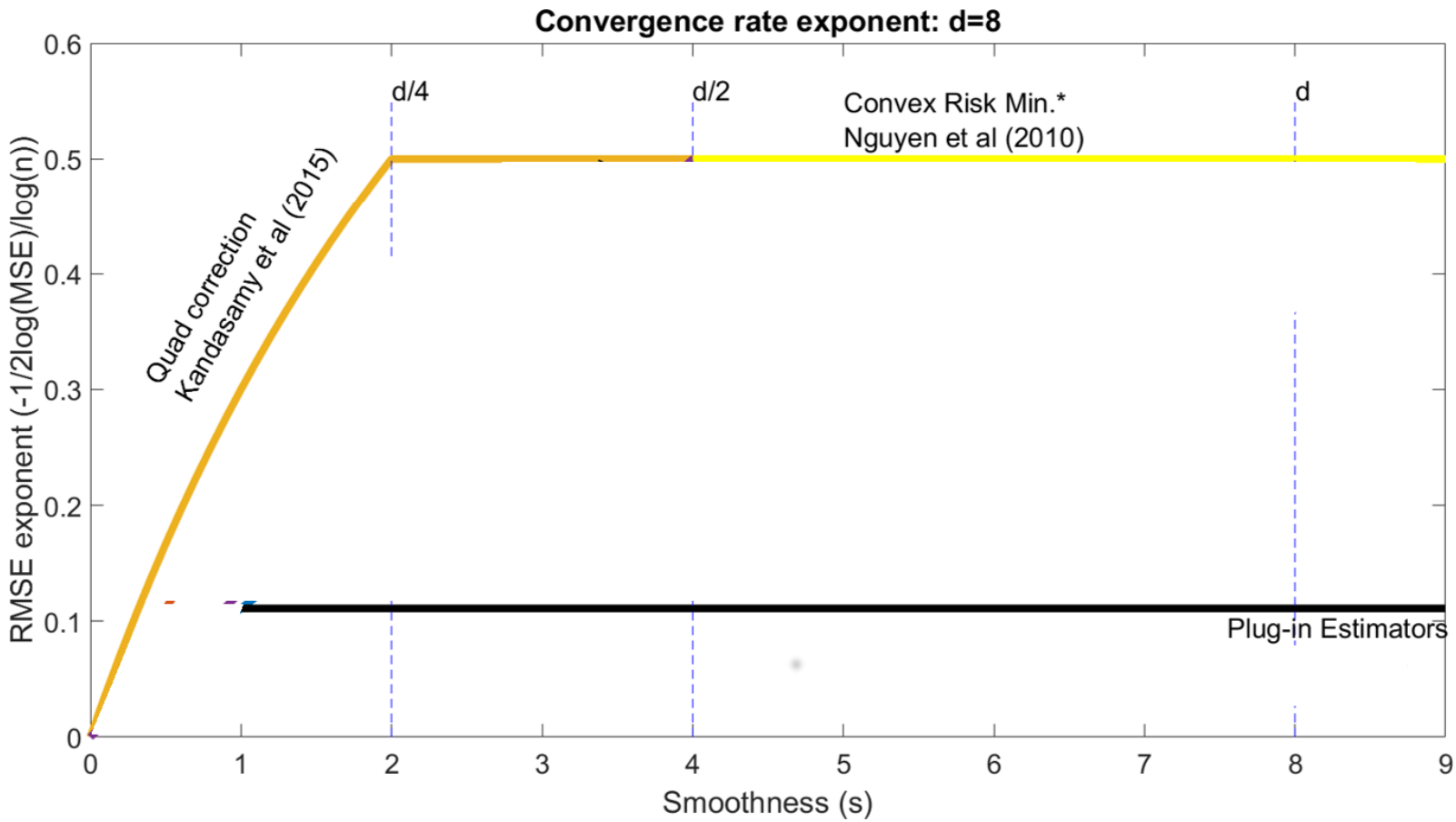


$\Rightarrow$ Kandasami's quadratic estimator achieves $O\left(\min\{n^{-3s/(2s+d)}, n^{-1/2}\}\right)$

[1] A. Krishnamurthy, K. Kandasamy, B. Póczos. Nonparametric estimation of Rényi divergence and friends, NIPS 2014.

[2] X. Nguyen, M. Wainwright, M. Jordan "Estimating divergence functionals and the likelihood ratio by convex risk minimization." IEEE Trans on Information Theory, 2010.

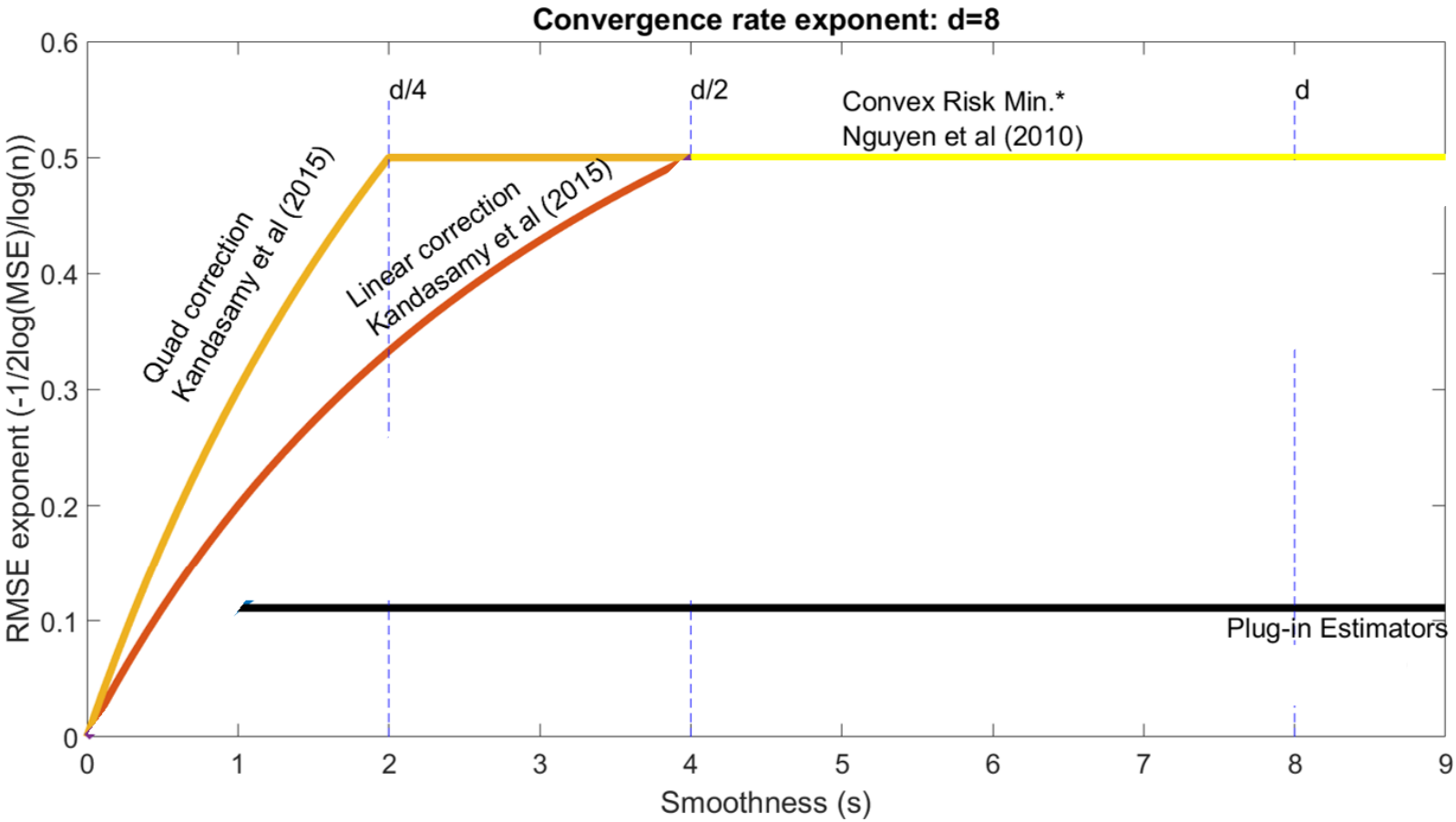
RMSE convergence rate comparisons[1]



$\Rightarrow$ Density plug-in is a weak learner: RMSE rate is $O(n^{-1/d})$

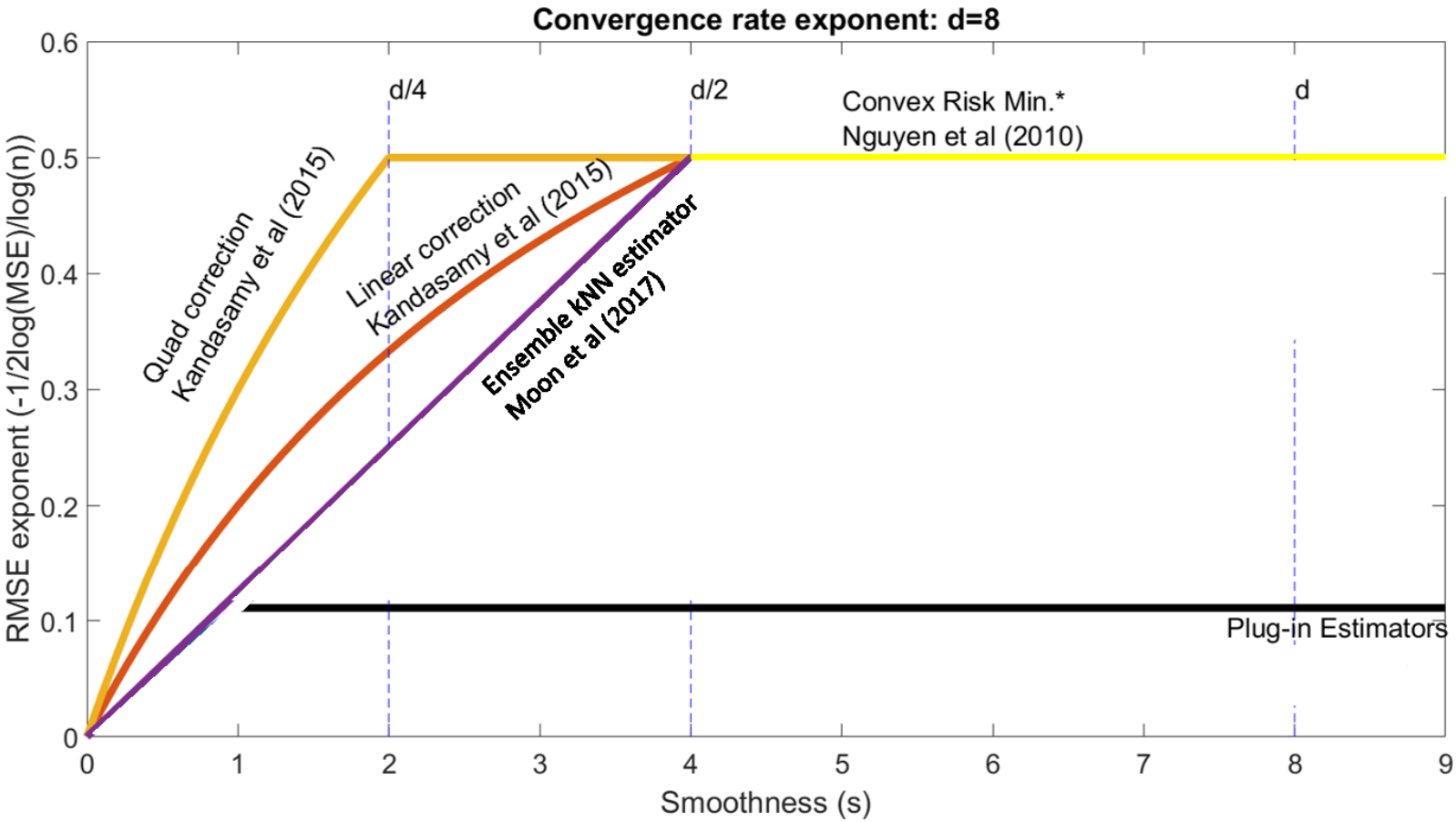
[1] K Moon, K Sricharan, K Greenewald, A. Hero. Ensemble Estimation of Information Divergence," Entropy, vol. 20, no. 8, p. 560, July 2018.

RMSE convergence rate comparisons[1]



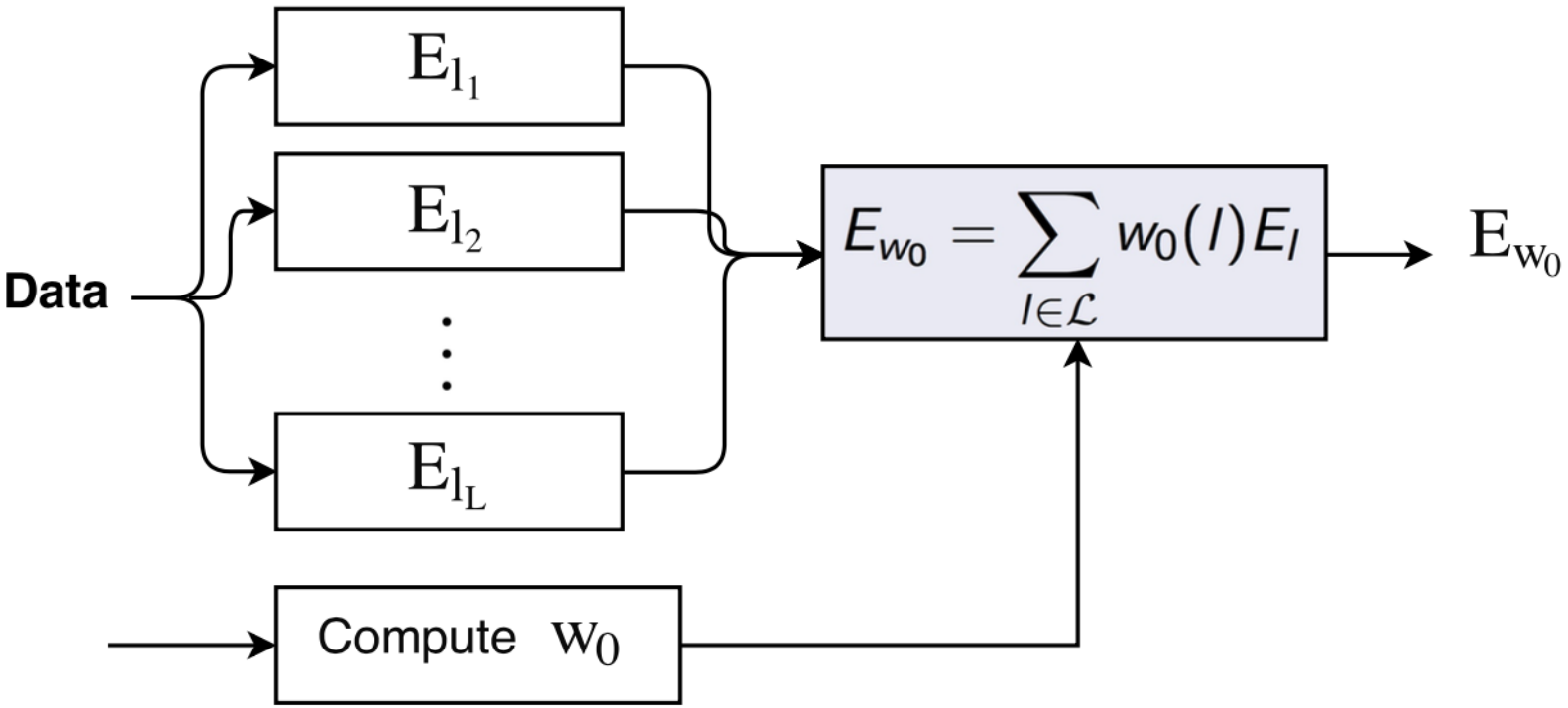
[1] A. Krishnamurthy, K. Kandasamy, B. Póczos. Nonparametric estimation of Rényi divergence and friends, NIPS 2014.

RMSE convergence rate comparisons[1]



[1] K Moon, K Sricharan, K Greenewald, A. Hero. Ensemble Estimation of Information Divergence," Entropy, vol. 20, no. 8, p. 560, July 2018.

Boosting ensembles concept



- $\{E_{l_i}\}_{i=1}^{L}$ ensemble of base estimators (weak learners)
- $\mathbf{w} = (w_0(l))_{l=1}^{L}$ a vector of boosting weights
- E_{w_0}: combined base estimators (boosted learner)

Choice of boosting weights

Approaches to weight selection

- Boosting classifiers with Adaboost[1] and other objective functions.
- Consensus clustering: unsupervised classification[2]
- Matching pursuit for dictionary learning[3]

[1] Y. Freund and R. E. Schapire (1996). Experiments with a new boosting algorithm. Intl Conf on Machine Learning. pp. 148-156.

[2] L. Tao, C. Ding and M. I. Jordan. "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization." Intl Conf on Data Mining, 2007, pp. 577-582.

[3] S. Mallat and Z. Zhang (1993). Matching pursuit with time frequency dictionaries. IEEE T. on Signal Processing 41:3397-3415

[4] Bickel, P. J., Ritov, Y. A., and Zakai, A. (2006). Some theory for generalized boosting algorithms. J. of Machine Learning Research, 705-732.

Choice of boosting weights

Approaches to weight selection

- Boosting classifiers with Adaboost[1] and other objective functions.
- Consensus clustering: unsupervised classification[2]
- Matching pursuit for dictionary learning[3]

Under some conditions such methods achieve Bayes optimal performance[4]

[1] Y. Freund and R. E. Schapire (1996). Experiments with a new boosting algorithm. Intl Conf on Machine Learning. pp. 148-156.

[2] L. Tao, C. Ding and M. I. Jordan. "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization." Intl Conf on Data Mining, 2007, pp. 577-582.

[3] S. Mallat and Z. Zhang (1993). Matching pursuit with time frequency dictionaries. IEEE T. on Signal Processing 41:3397-3415

[4] Bickel, P. J., Ritov, Y. A., and Zakai, A. (2006). Some theory for generalized boosting algorithms. J. of Machine Learning Research, 705-732.

Choice of boosting weights

Approaches to weight selection

- Boosting classifiers with Adaboost[1] and other objective functions.
- Consensus clustering: unsupervised classification[2]
- Matching pursuit for dictionary learning[3]

Under some conditions such methods achieve Bayes optimal performance[4]

In each of these cases, the ensemble weights cannot be computed offline. They are data dependent.

---

[1] Y. Freund and R. E. Schapire (1996). Experiments with a new boosting algorithm. Intl Conf on Machine Learning. pp. 148-156.

[2] L. Tao, C. Ding and M. I. Jordan. "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization." Intl Conf on Data Mining, 2007, pp. 577-582.

[3] S. Mallat and Z. Zhang (1993). Matching pursuit with time frequency dictionaries. IEEE T. on Signal Processing 41:3397-3415

[4] Bickel, P. J., Ritov, Y. A., and Zakai, A. (2006). Some theory for generalized boosting algorithms. J. of Machine Learning Research, 705-732.

Choice of boosting weights

Approaches to weight selection

- Boosting classifiers with Adaboost[1] and other objective functions.
- Consensus clustering: unsupervised classification[2]
- Matching pursuit for dictionary learning[3]

Under some conditions such methods achieve Bayes optimal performance[4]

In each of these cases, the ensemble weights cannot be computed offline. They are data dependent.

$\Rightarrow$ Our proposed ensemble divergence estimator has rate-optimal weights that are not data dependent

---

[1] Y. Freund and R. E. Schapire (1996). Experiments with a new boosting algorithm. Intl Conf on Machine Learning. pp. 148-156.

[2] L. Tao, C. Ding and M. I. Jordan. "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization." Intl Conf on Data Mining, 2007, pp. 577-582.

[3] S. Mallat and Z. Zhang (1993). Matching pursuit with time frequency dictionaries. IEEE T. on Signal Processing 41:3397-3415

[4] Bickel, P. J., Ritov, Y. A., and Zakai, A. (2006). Some theory for generalized boosting algorithms. J. of Machine Learning Research, 705-732.

Back to information divergence estimation problem

Information divergence function

$$D_g(f_1 \| f_2) = \int g\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x) dx$$

where $g(u)$ is a convex function on $\mathbb{R}^+$ and $g(1) = 0$.

[1] M. Noshad and A. Hero (2018). Scalable Hash-Based Estimation of Divergence Measures. AISTAT.

[2] K Moon and A Hero (2014). Multivariate f-divergence estimation with confidence. NIPS, pp. 2420-2428.

[3] Sricharan, K., Wei, D., and Hero, A. (2013). Ensemble estimators for multivariate entropy estimation. IEEE transactions on information theory, 59(7), 4374-4388.

Back to information divergence estimation problem

Information divergence function

$$D_g(f_1 \| f_2) = \int g\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x) dx$$

where $g(u)$ is a convex function on $\mathbb{R}^+$ and $g(1) = 0$.

Proposed base estimator: Locality Sensitive Hashing (LSH) plug-in estimator[1]

[1] M. Noshad and A. Hero (2018). Scalable Hash-Based Estimation of Divergence Measures. AISTAT.

[2] K Moon and A Hero (2014). Multivariate f-divergence estimation with confidence. NIPS, pp. 2420-2428.

[3] Sricharan, K., Wei, D., and Hero, A. (2013). Ensemble estimators for multivariate entropy estimation. IEEE transactions on information theory, 59(7), 4374-4388.

Back to information divergence estimation problem

Information divergence function

$$D_g(f_1 \| f_2) = \int g\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x) dx$$

where $g(u)$ is a convex function on $\mathbb{R}^+$ and $g(1) = 0$.

Proposed base estimator: Locality Sensitive Hashing (LSH) plug-in estimator[1]

Assumptions[2] (similar to kernel density plug-in entropy estimator [3])

- f_1 and f_2 have common bounded support set $\Omega \subset \mathbb{R}^d$.
- $0 < C_L \leq f_1, f_2 \leq C_U < \infty$ on Ω.
- Densities f_1 and f_2 are both differentiable of order d.
- g is Lipschitz continuous.

---

[1] M. Noshad and A. Hero (2018). Scalable Hash-Based Estimation of Divergence Measures. AISTAT.

[2] K Moon and A Hero (2014). Multivariate f-divergence estimation with confidence. NIPS, pp. 2420-2428.

[3] Sricharan, K., Wei, D., and Hero, A. (2013). Ensemble estimators for multivariate entropy estimation. IEEE transactions on information theory, 59(7), 4374-4388.

Locality sensitive hashing plug-in estimator

Define numbers of samples N and M in class 1 and class 2, respectively,

$$N = \sum_{i=1}^{n} Y_i, \qquad M = \sum_{i=1}^{n} (1 - Y_i) = n - N$$

Locality sensitive hashing plug-in estimator

Define numbers of samples N and M in class 1 and class 2, respectively,

$$N = \sum_{i=1}^{n} Y_i, \qquad M = \sum_{i=1}^{n} (1 - Y_i) = n - N$$

LSH-based divergence estimator

- Step 1: Apply LSH quantizer $\{Q_i(x)\}_{i=1}^{F}$ to bin X_i into F buckets

$$N_i = \sum_{j=1}^{n} Y_j Q_i(X_j), \qquad M_i = \sum_{j=1}^{n} (1 - Y_j) Q_i(X_j),$$

- Step 2: Compute divergence plug-in estimator using histograms $\{(N_i, M_i)\}$

$$\widehat{D}_g(f_1 \| f_2) := \sum_{i: M_i > 0} g\left(\frac{N_i/N}{M_i/M}\right) M_i/M$$

Locality sensitive hashing plug-in estimator

$$\widehat{D}_g(f_1 \| f_2) := \sum_{i:M_i>0} g\left(\frac{N_i/N}{M_i/M}\right) M_i/M$$

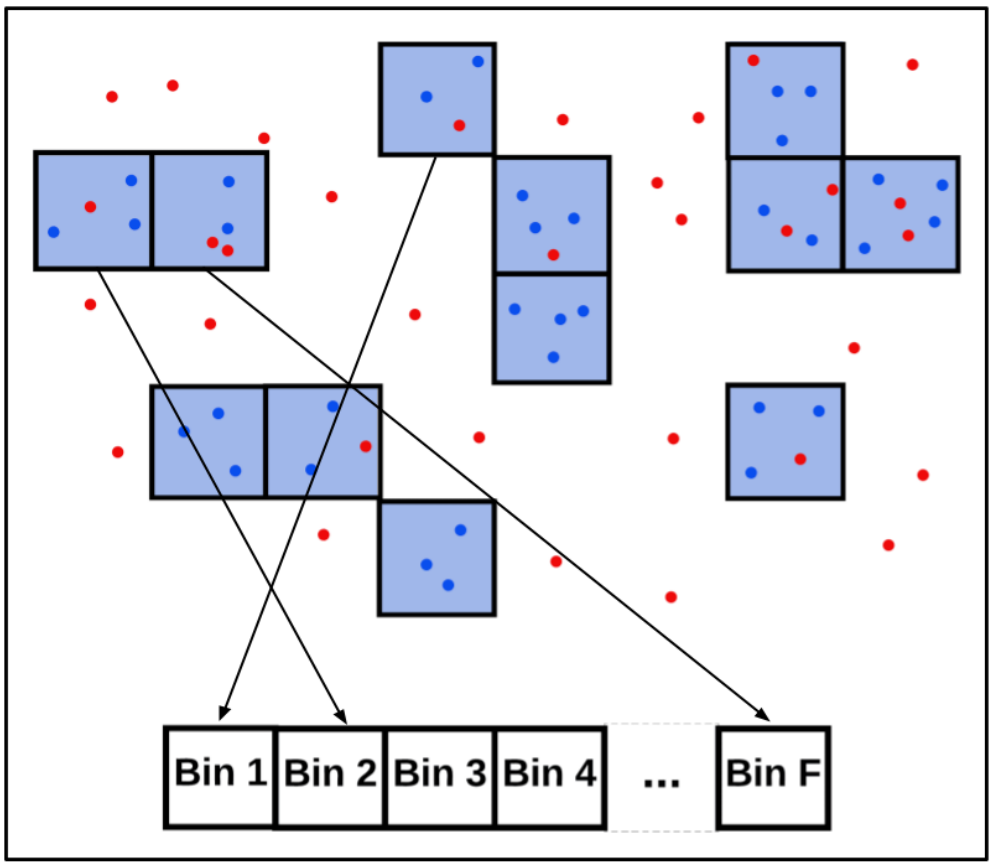


Figure: LSH quantizes **X** data with random displacement b and cell resolution ϵ

Locality sensitive hashing plug-in estimator: bias and variance

Theorem (Bias Expansion)

If f_1 and f_2 are d-times differentiable, the mean of $\widehat{D}_g$ is

$$\mathbb{E}[\widehat{D}_g] = D(f_1 \| f_2) + \mathbb{B}(\widehat{D}_g)$$

$$\mathbb{B}(\widehat{D}_g) = \sum_{i=1}^{d} C_i \epsilon^i + O\left(\frac{1}{n\epsilon^d}\right).$$

Locality sensitive hashing plug-in estimator: bias and variance

Theorem (Bias Expansion)

If f_1 and f_2 are d-times differentiable, the mean of $\widehat{D}_g$ is

$$\mathbb{E}[\widehat{D}_g] = D(f_1 \| f_2) + \mathbb{B}(\widehat{D}_g)$$

$$\mathbb{B}(\widehat{D}_g) = \sum_{i=1}^{d} C_i \epsilon^i + O\left(\frac{1}{n\epsilon^d}\right).$$

Theorem (Variance)

The variance of the hash-based estimator can be bounded as

$$\mathbb{V}(\widehat{D}_g) \leq O\left(\frac{1}{n}\right).$$

Locality sensitive hashing plug-in estimator: bias and variance

Theorem (Bias Expansion)

If f_1 and f_2 are d-times differentiable, the mean of $\widehat{D}_g$ is

$$\mathbb{E}[\widehat{D}_g] = D(f_1 \| f_2) + \mathbb{B}(\widehat{D}_g)$$

$$\mathbb{B}(\widehat{D}_g) = \sum_{i=1}^{d} C_i \epsilon^i + O\left(\frac{1}{n\epsilon^d}\right).$$

Theorem (Variance)

The variance of the hash-based estimator can be bounded as

$$\mathbb{V}(\widehat{D}_g) \leq O\left(\frac{1}{n}\right).$$

$\Rightarrow$ Choosing $\epsilon = O\left(n^{-1/2d}\right)$ forces bias remainder to $O\left(\frac{1}{n\epsilon^d}\right) = O(1/\sqrt{n})$

Locality sensitive hashing plug-in estimator: bias and variance

Theorem (Bias Expansion)

If f_1 and f_2 are d-times differentiable, the mean of $\widehat{D}_g$ is

$$\mathbb{E}[\widehat{D}_g] = D(f_1 \| f_2) + \mathbb{B}(\widehat{D}_g)$$

$$\mathbb{B}(\widehat{D}_g) = \sum_{i=1}^{d} C_i \epsilon^i + O\left(\frac{1}{n\epsilon^d}\right).$$

Theorem (Variance)

The variance of the hash-based estimator can be bounded as

$$\mathbb{V}(\widehat{D}_g) \leq O\left(\frac{1}{n}\right).$$

$\Rightarrow$ Choosing $\epsilon = O\left(n^{-1/2d}\right)$ forces bias remainder to $O\left(\frac{1}{n\epsilon^d}\right) = O(1/\sqrt{n})$

$\Rightarrow$ This makes the slowest term in the bias decay as $\mathbb{B}(\widehat{D}_g) = O(n^{-1/2d})$

Ensemble bias reduction

- Let $\{\widehat{D}_g^{\epsilon(t)}\}_{t\in\mathcal{L}}$ be a set of $L = |\mathcal{L}|$ base learners.
- $\mathcal{L} := \{t_1, ..., t_L\}$ is a set of index values.
- $\epsilon(t) = tn^{-1/2d}$ is a set of histogram bandwidth parameters.

Define: Ensemble divergence estimator $\widehat{D}_w := \sum_{t\in\mathcal{L}} w(t)\widehat{D}_{\epsilon(t)}$

Ensemble bias reduction

- Let $\{\widehat{D}_g^{\epsilon(t)}\}_{t\in\mathcal{L}}$ be a set of $L = |\mathcal{L}|$ base learners.
- $\mathcal{L} := \{t_1, ..., t_L\}$ is a set of index values.
- $\epsilon(t) = tn^{-1/2d}$ is a set of histogram bandwidth parameters.

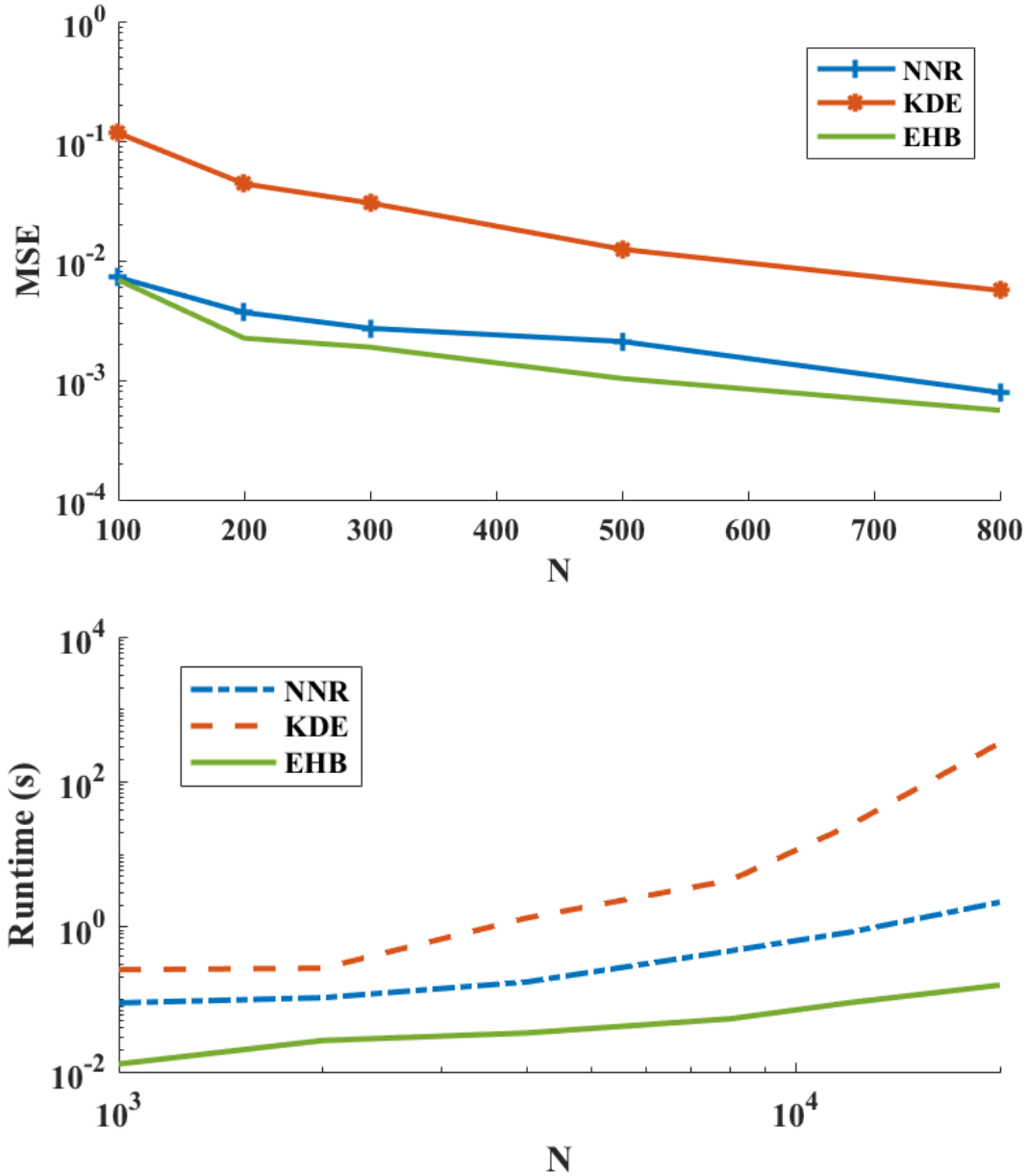
Define: Ensemble divergence estimator $\widehat{D}_w := \sum_{t\in\mathcal{L}} w(t)\widehat{D}_{\epsilon(t)}$

$$\mathbb{B}\left[\widehat{D}_w\right] = \sum_{i=1}^{d} C_i n^{-i/2d} \sum_{t=1}^{d} w(t)t^i + O\left(\frac{1}{\sqrt{n}}\right)$$

$\Rightarrow$ Bias becomes $O(1/\sqrt{n})$ if $w(t)$ is selected (offline) according to

$$\begin{aligned} \min_{w} \quad & \|w\|_2 \\ \textit{subject to} \quad & \sum_{t\in\mathcal{L}} w(t) = 1, \\ & \sum_{t\in\mathcal{L}} w(t)t^i = 0, i \in \mathbb{N}, i \leq d \end{aligned}$$

Simulation validation: Hellinger divergence for 2D shifted normals.



Benchmark learner vs HP bound learner

Simulation: classification of 2 mean shifted 10 dim Gaussian densities

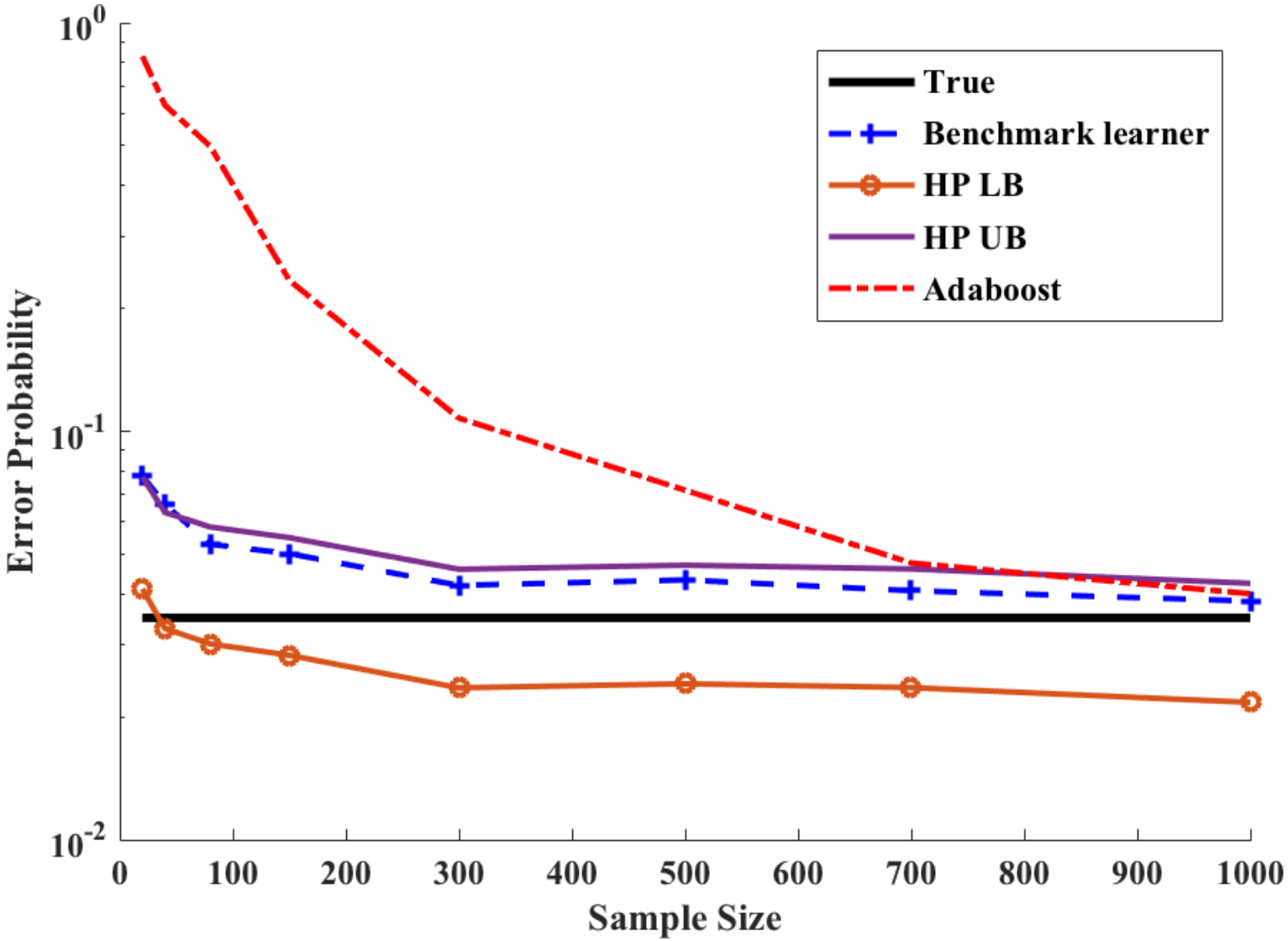


Figure: Note: ensemble benchmark learner is within envelope of HP bound learner.

Benchmark learner for multiclass classification

Gaussian $\mathcal{N}_2(\mu, \mathbf{I}_2)$, $K = 4$ classes $\mu_1, \ldots, \mu_4$

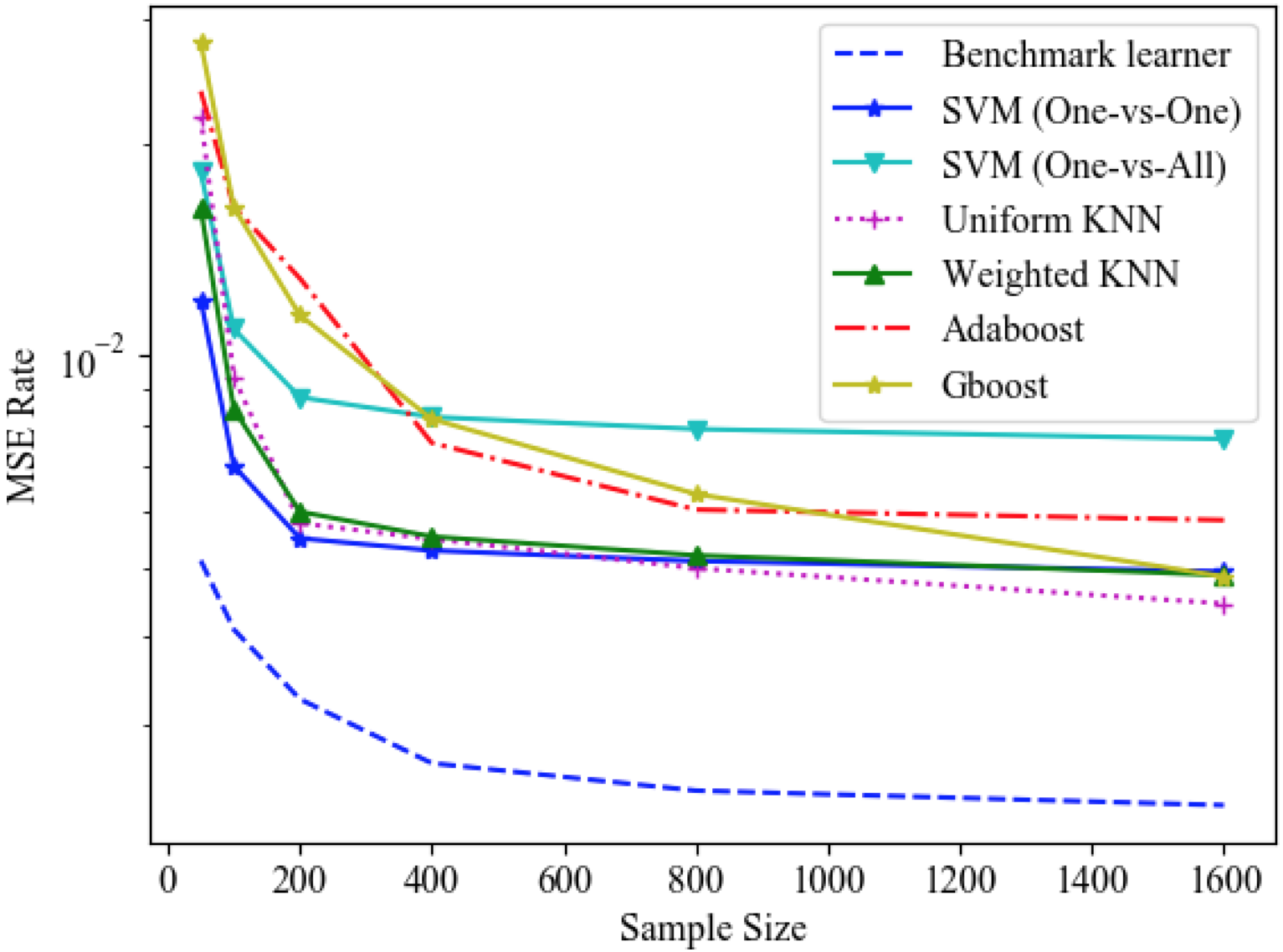
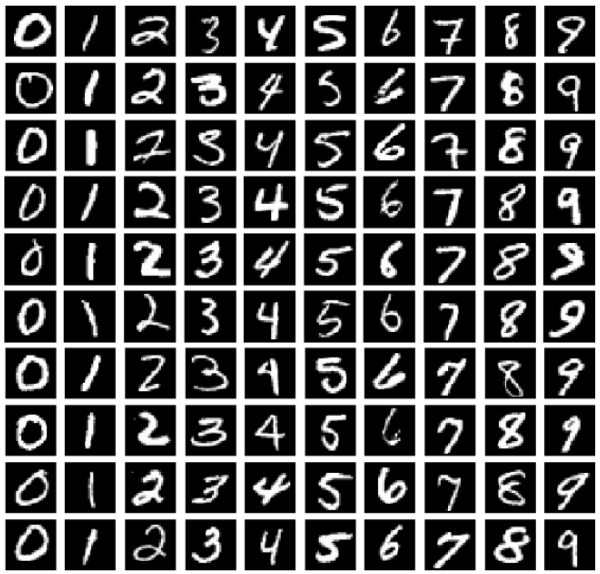
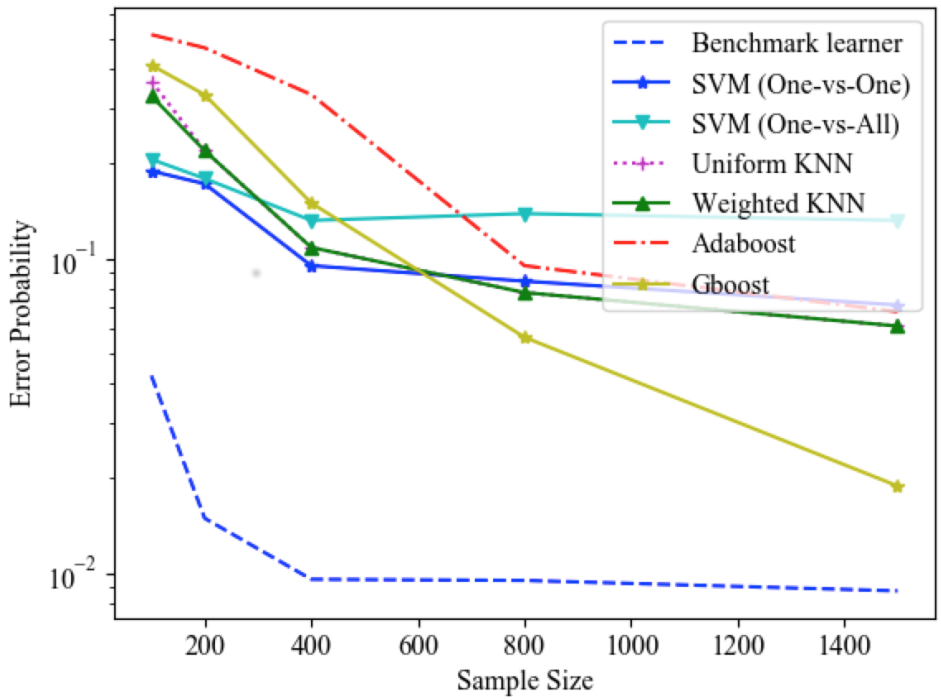


Figure: Note: Benchmark learner attains much lower MSE in estimating Bayes error

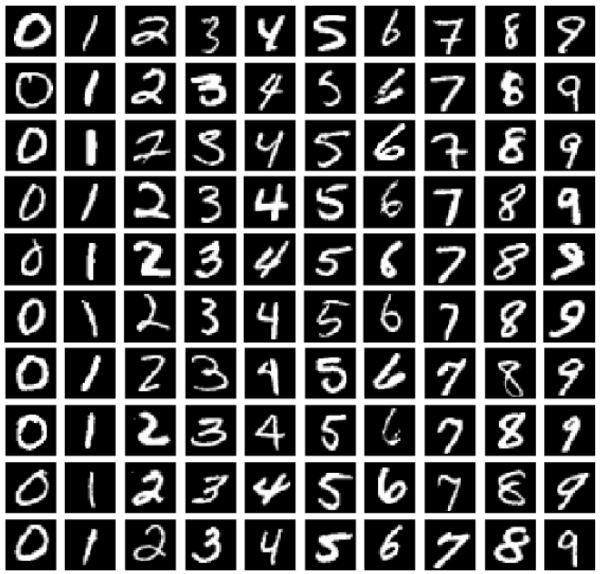
Benchmarking MNIST digit classification

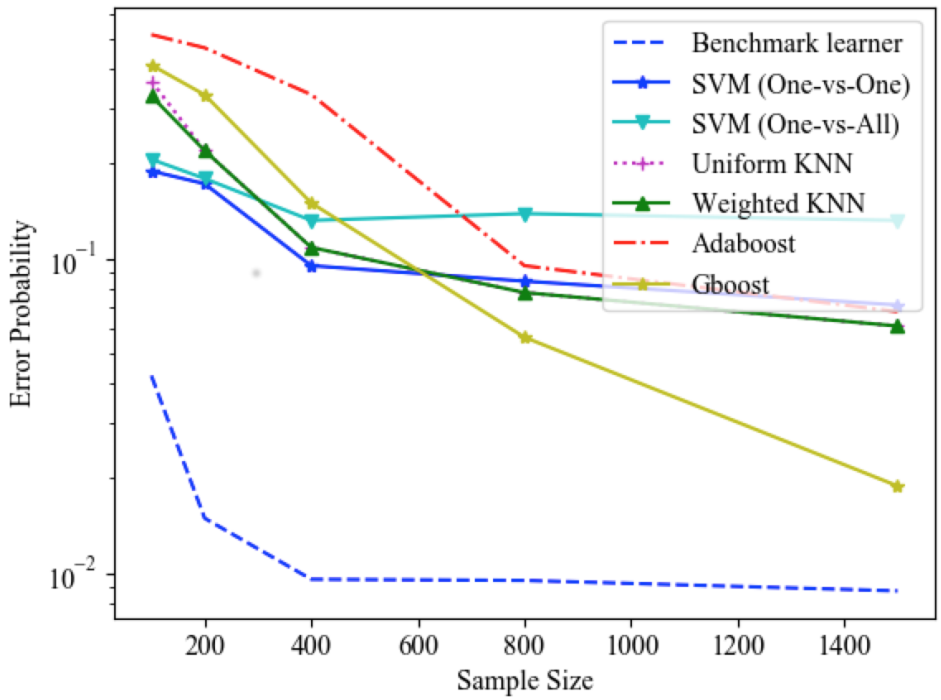




- MNIST handwritten digits with $K = 10$ classes
- Digit image data has $d = 64$ dimensions
- $n = 1795$ samples

Benchmarking MNIST digit classification





- MNIST handwritten digits with $K = 10$ classes
- Digit image data has $d = 64$ dimensions
- $n = 1795$ samples

Note: Gboost comes closest to benchmark but there is large margin for improvement.

Action classification for navigating SCITOS G5 robot

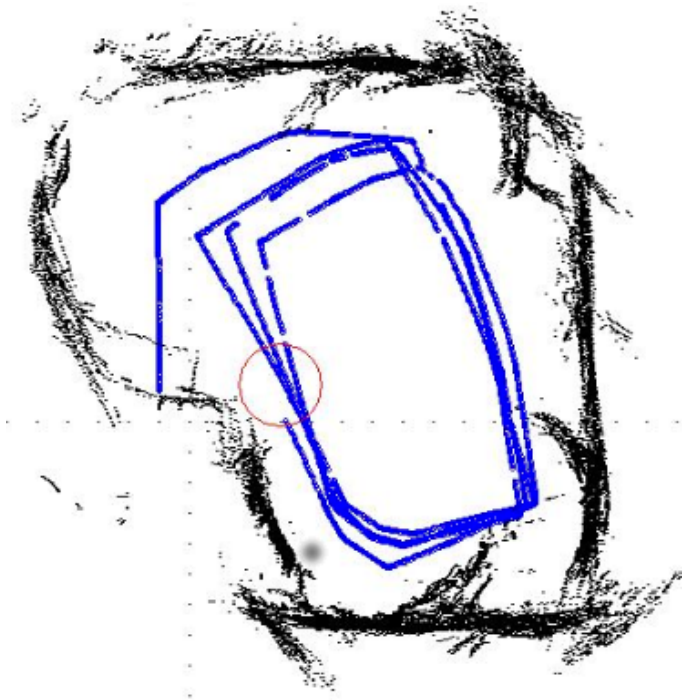


Figure: SCITOS G5 w/ 24 sensors

Robot's navigation path

- SCITOS robot navigation data from UCI database
- 4 robot control actions: Sharp R, Slight R, Forward, Slight L
- 24 equispaced ultrasound sensors on robot skirt
- $n = 5456$ action instances collected in experiment

Action classification for navigating SCITOS G5 robot

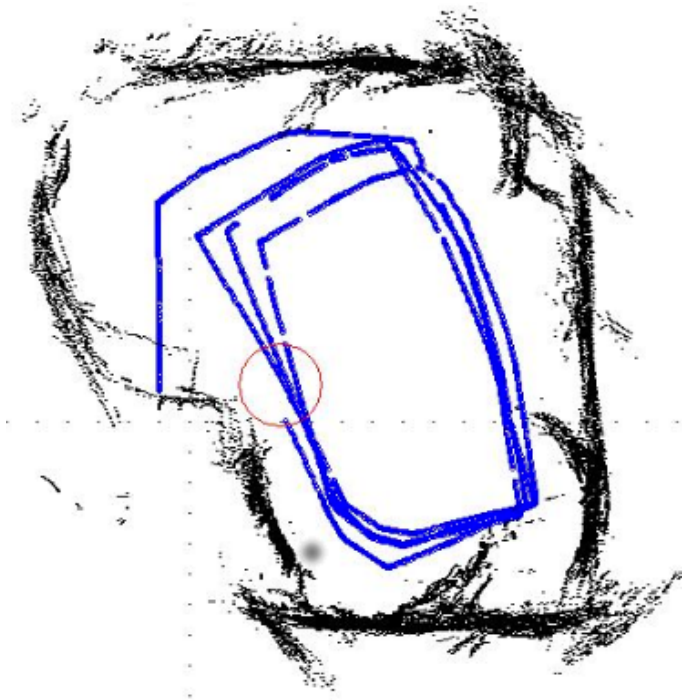


Figure: SCITOS G5 w/ 24 sensors Robot's navigation path

- SCITOS robot navigation data from UCI database
- 4 robot control actions: Sharp R, Slight R, Forward, Slight L
- 24 equispaced ultrasound sensors on robot skirt
- $n = 5456$ action instances collected in experiment

Q. Without knowing control law, how well can we learn to predict the actions from the sensor readings?

Action classification: DoE w/ navigating SCITOS G5 robot

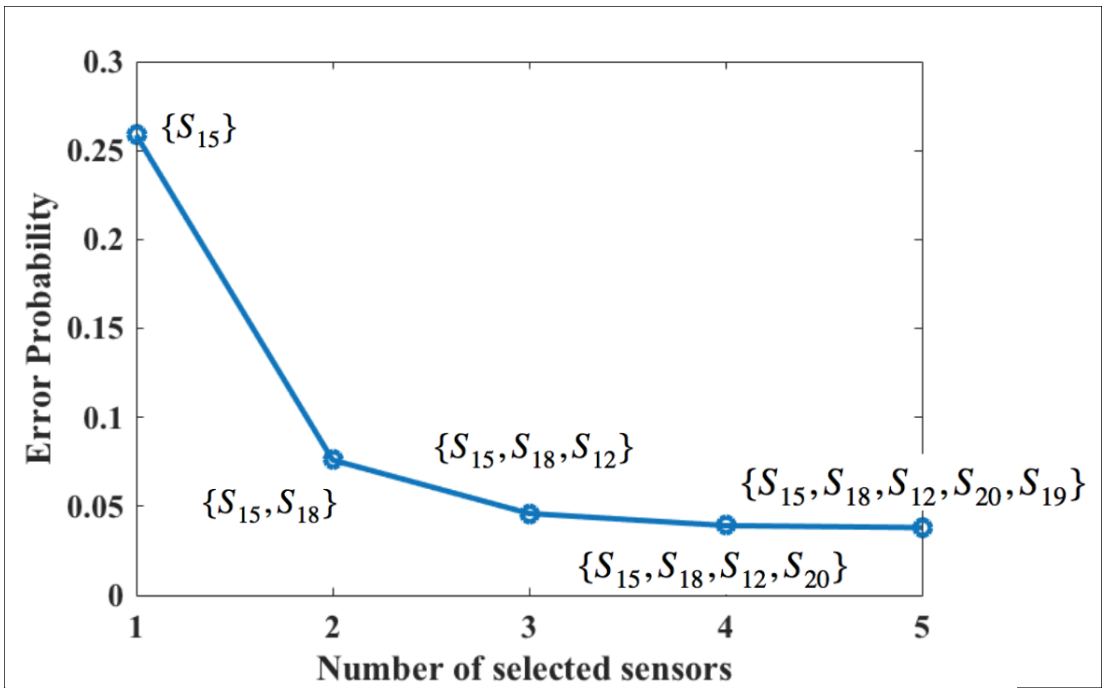
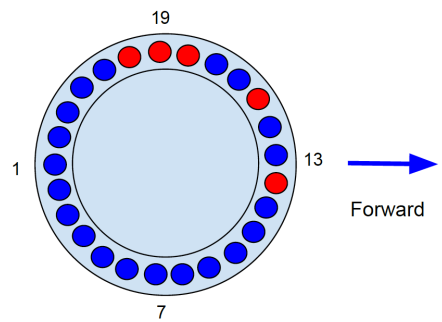


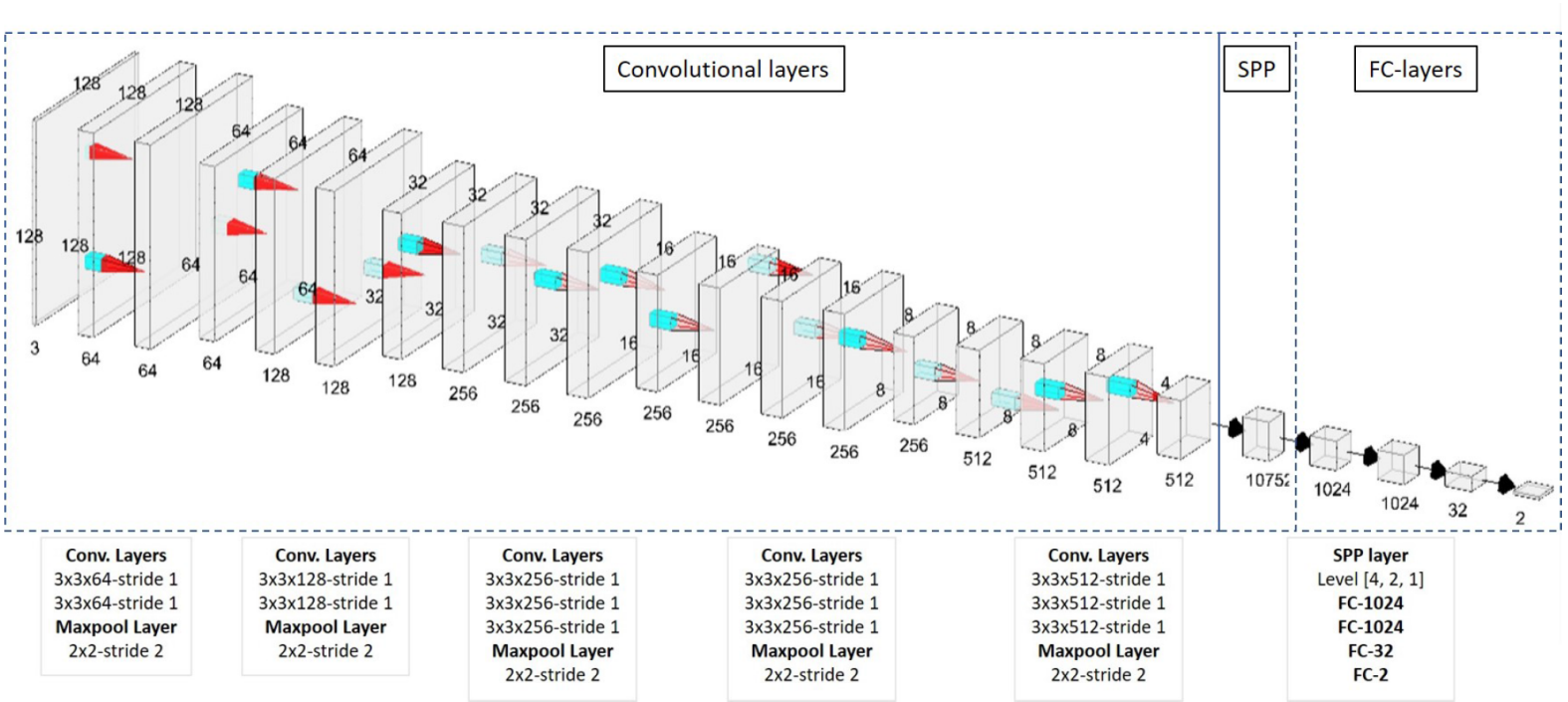
Figure: Sensor-selection curve



The top ranked 5 sensors selected

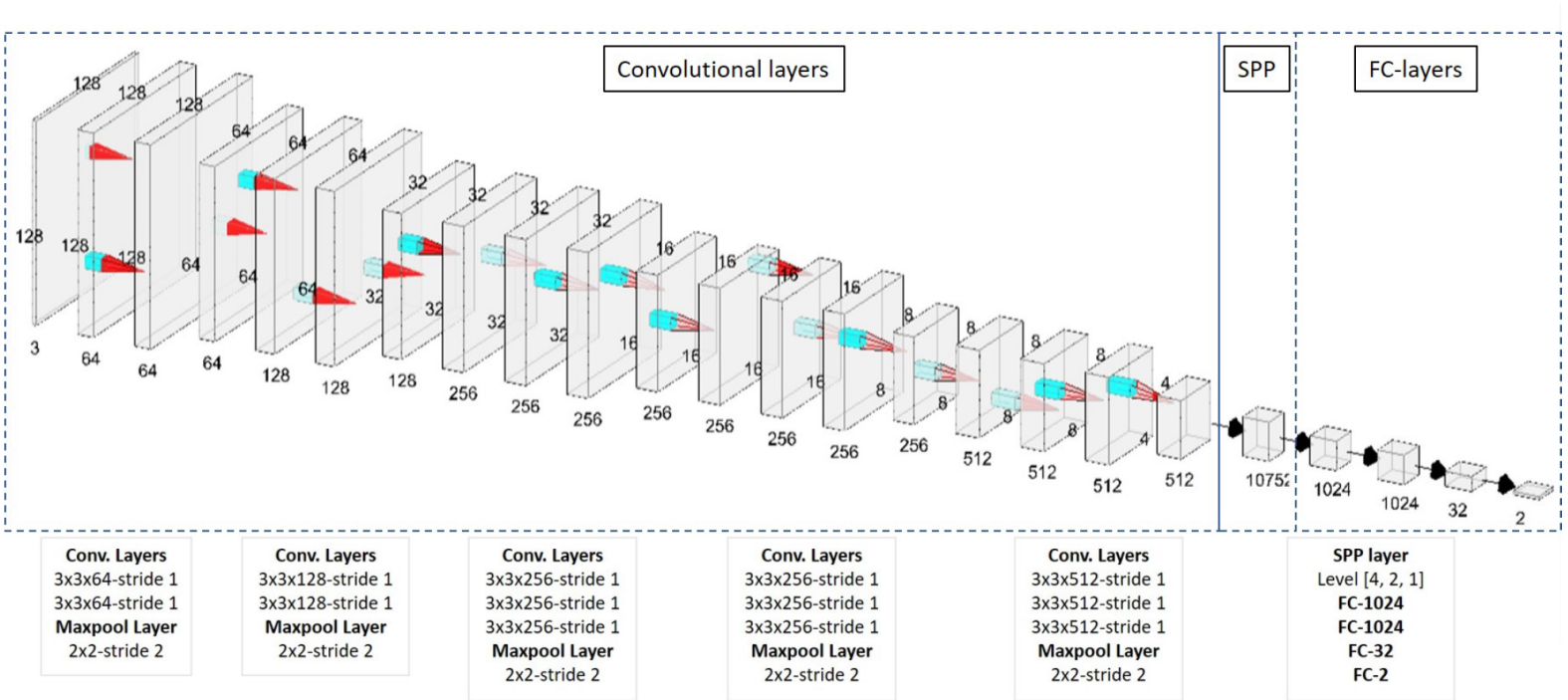
- 4 HP divergences estimated for each of 4 one-vs-all-classifiers
- These estimates mapped to average Bayes error using upper HP bound
- Greedy step-up and step down sensor selection procedures implemented
- 3 sensors suffice to force estimated misclassification error below 0.05.

Mutual information estimation: application to DNN information bottleneck



Convolutional neural network (CNN) for image classification[1]

Mutual information estimation: application to DNN information bottleneck



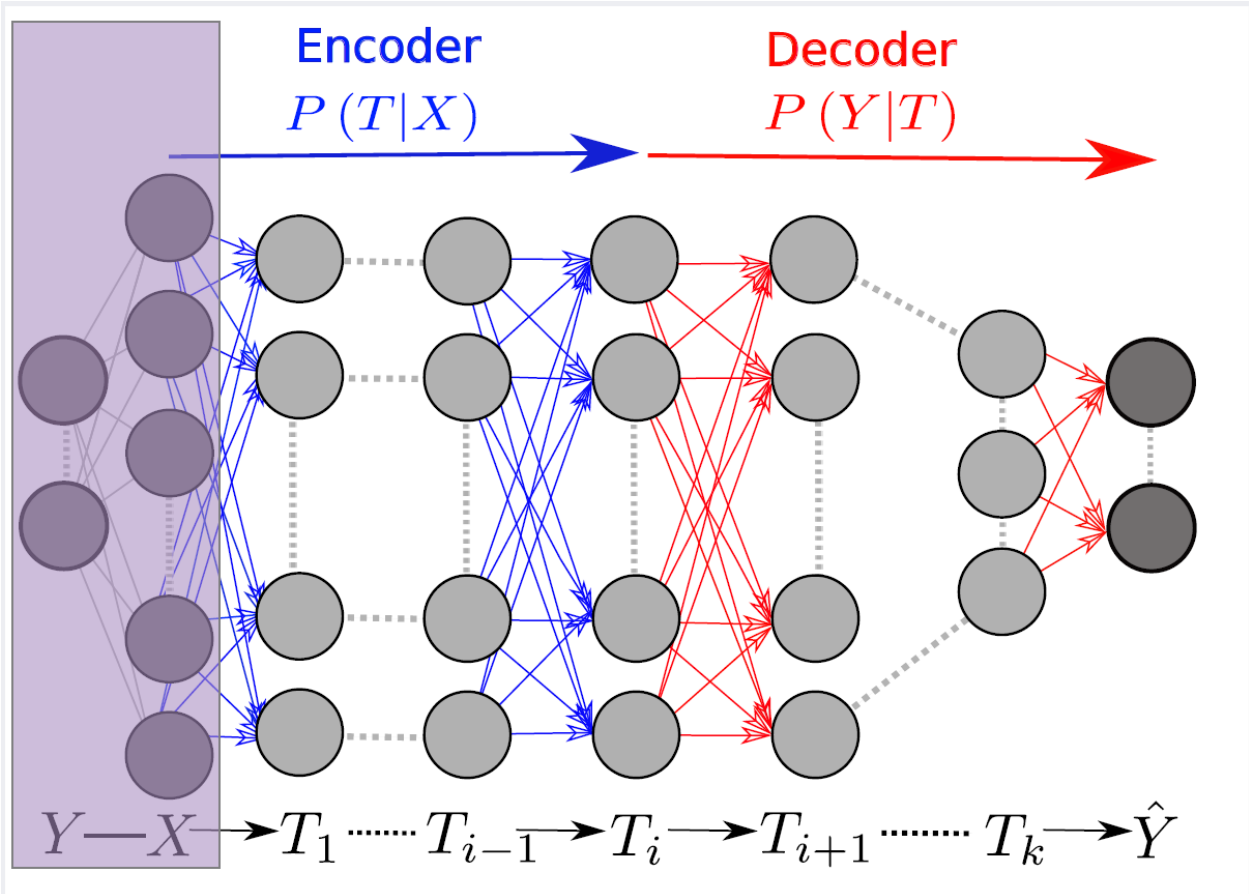
Convolutional neural network (CNN) for image classification[1]

DNN design questions:

- How many input (convolution) layers?
- How many output (discrimination) layers?
- How much pooling and what kind of activation functions?
- How to set learning rate parameters?

[1] B. DuFumier. A new deep learning approach to solar flare prediction. ENSTA internship report, Sept. 2018

Tishby's framework: encoder/decoder information bottleneck

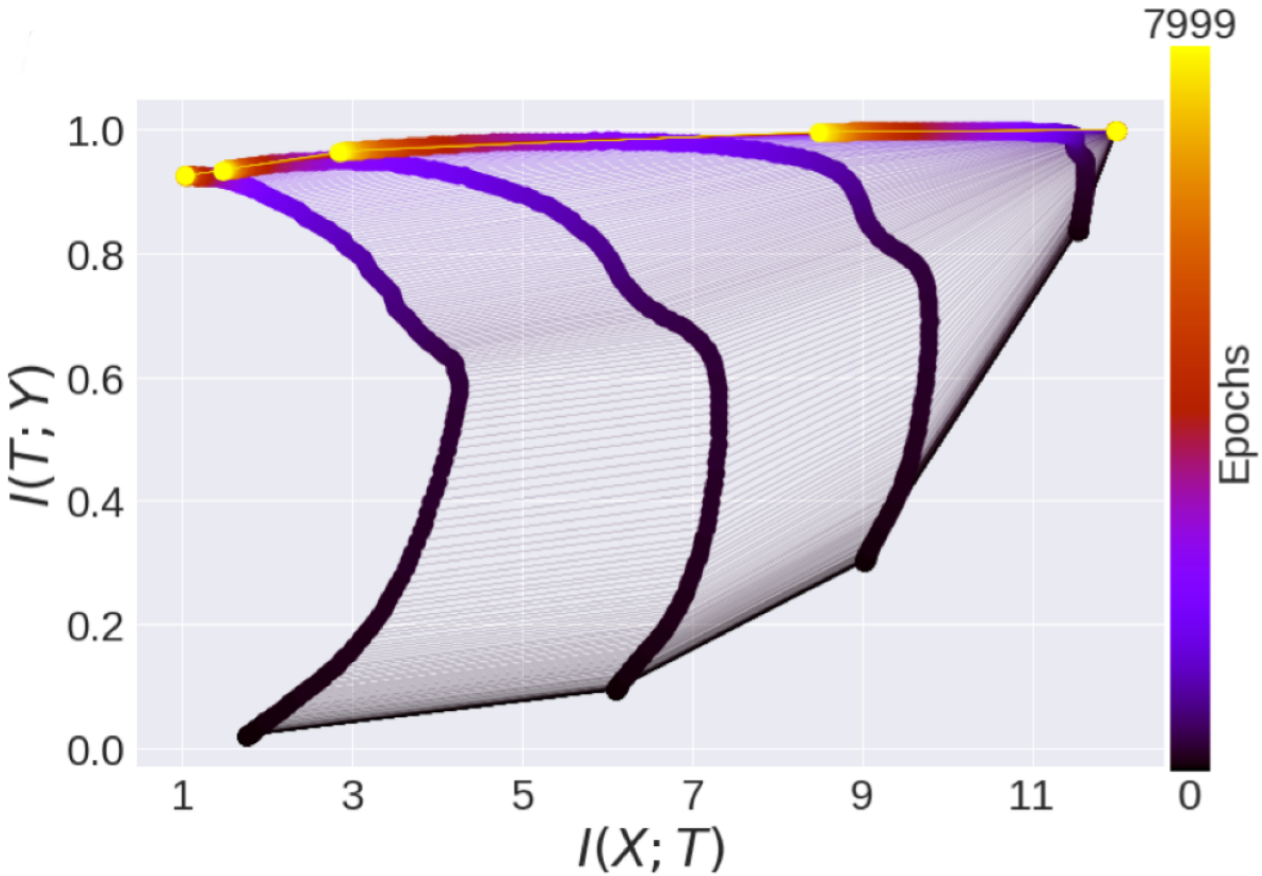


- Encoder I/O: input X, ouptut T (features)
- Decoder I/O: input T, output Y (labels)

[1] R Schwartz-Ziv and N Tishby. Opening the black box of deep neural networks via information. arXiv 2017

[2] AM Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, BD Tracey, DD. Cox, "On the information bottleneck theory of deep learning," ICLR 2018

The information plane



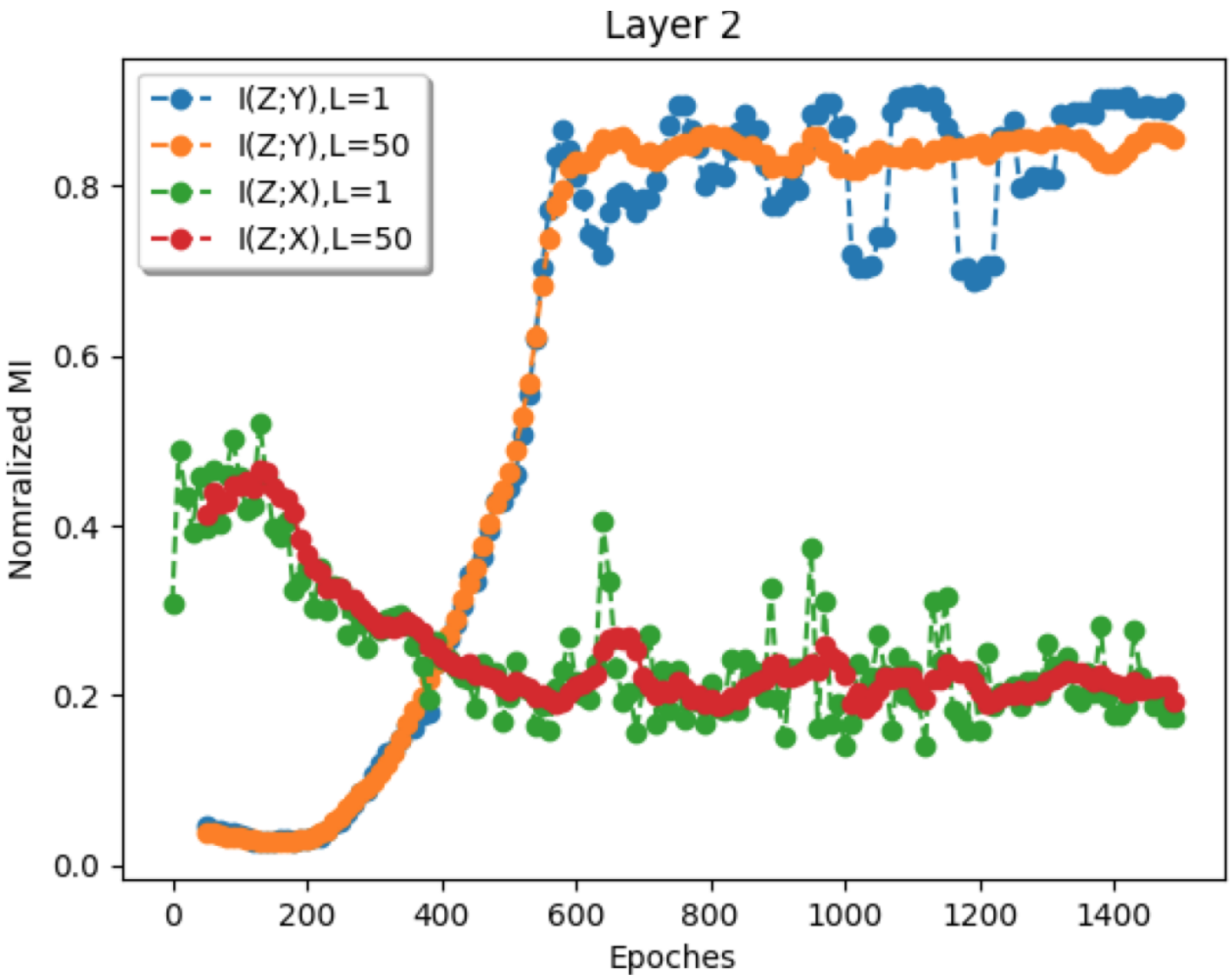
- Plot of training-trajectories of $[I(X;T_i), I(T_i;Y)]$ for different layers T_i

$$I(X;T) = \int f_{XT} \log\left(\frac{f_{XT}}{f_X f_T}\right), \quad I(T;Y) = \int f_{TY}\left(\frac{f_{TY}}{f_T f_Y}\right)$$

[1] R Schwartz-Ziv and N Tishby. Opening the black box of deep neural networks via information. arXiv 2017

[2] AM Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, BD Tracey, DD. Cox, "On the information bottleneck theory of deep

Information plane improvements via ensemble MI estimation



- 10-8-6-4-2 ReLu MLP trained on 10,000 samples of 10D Gaussian
- MI with $L = 1$ (green&blue) is identical to Tishby's implementation
- Proposed ensemble MI implementation (red&orange) is more stable

Summary and open problems

Main takeaways

- Meta-learning v0: Constructing weak base learners of Bayes error rate (BER) for multiclass
- Meta-learning v1: Ensemble learning of BER with optimized data-independent weights
- LSH ensemble method achieves rate optimal performance in both computational complexity and sample complexity
- Illustrated for MNIST digit classification, feature selection, and DNN information bottleneck

Summary and open problems

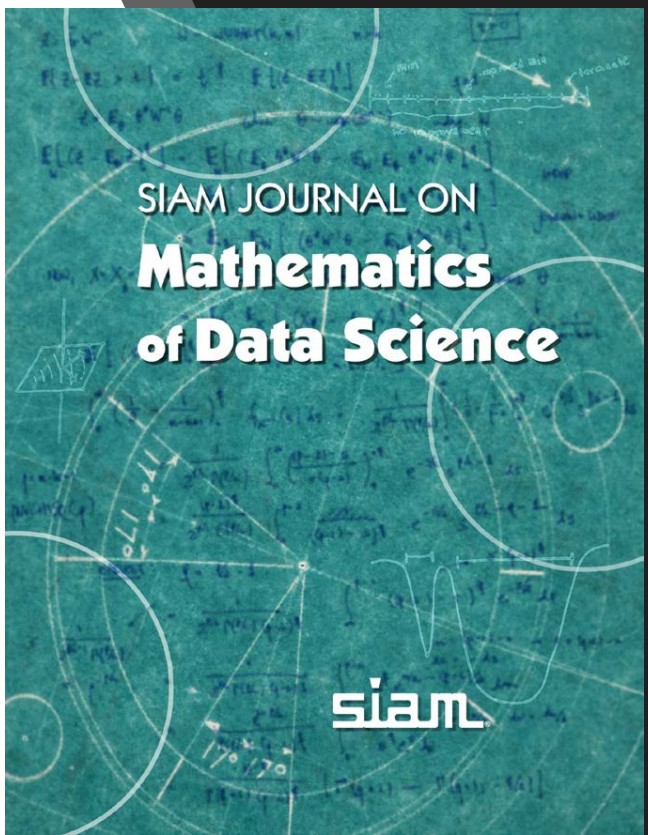
Main takeaways

- Meta-learning v0: Constructing weak base learners of Bayes error rate (BER) for multiclass
- Meta-learning v1: Ensemble learning of BER with optimized data-independent weights
- LSH ensemble method achieves rate optimal performance in both computational complexity and sample complexity
- Illustrated for MNIST digit classification, feature selection, and DNN information bottleneck

Open problems

- Weakening of d-times differentiability and strict boundedness on density
- High-dimensional learning regimes: $n, d \to \infty$
- Partially labeled data: Learning to benchmark semisupervised classification

New journal



- Publishes work that advances mathematical, statistical, and computational methods in the context of data and information sciences. We invite papers that present significant advances in this context, including applications to science, engineering, business, and medicine
- Editor-in-chief: Tamara G. Kolda (Sandia)
- Section Editors:
  - Alfred Hero (Michigan)
  - Michael I. Jordan (Berkeley)
  - Robert Nowak (Wisconsin)
  - Joel A. Tropp (Caltech)
- SIAM's newest journal will begin to take author submissions in Spring 2018
- http://www.siam.org/journals/simods.php

Thanks

My current group

Farshad Hirarichi (Post-doc)
Salimeh Sekeh (Post-doc)
Li Xu (Post-doc)

Joel LeBlanc (EECS)
Brandan Oselio (EECS)
Elizabeth Hou (EECS)
Hoanan Zhu (EECS)
Oskar Singer (EECS)
Neo Charalambides (EECS)
Yaya Zhai (PIBS)
Yun Wei (AIM)
Alex Zaitlef (AIM)
Audelia Wittbrodt (AP)
Byoung Jang (Statistics)

Academic collaborators

Indika Rajapakse (UM DCMB)
Angela Violi (UM ME)
Sara Pozzi (UM NERS)
Vahid Tarokh (Duke)
Geoff Ginsburg (Duke)
Yasin Yilmaz (USF)
Taposh Banerjee (Harvard)
Yoann Altman (Heriot-Watt)
Stephen McGlaughlin (Heriot-Watt)
Alex Jung (Aalto)

Visiting students

Hafiz Tiomoko (SupElec)
Gert van den Berg (Erasmus)

Other Collaborators

Pin Yu Chen (IBM)
Sijia Liu (IBM)
Sung Jin Hwang (Google)

Brian Sadler (ARL)
Ed Zelnio (AFRL)
Sutunay Choudhouri (PNNL)
Earl Lawrence (LANL)

Current sponsors

DOE NNSA
DARPA Prometheus
DARPA DeepPurple
ARO MURI Programs
AFRL ATR Center
Xerox Parc