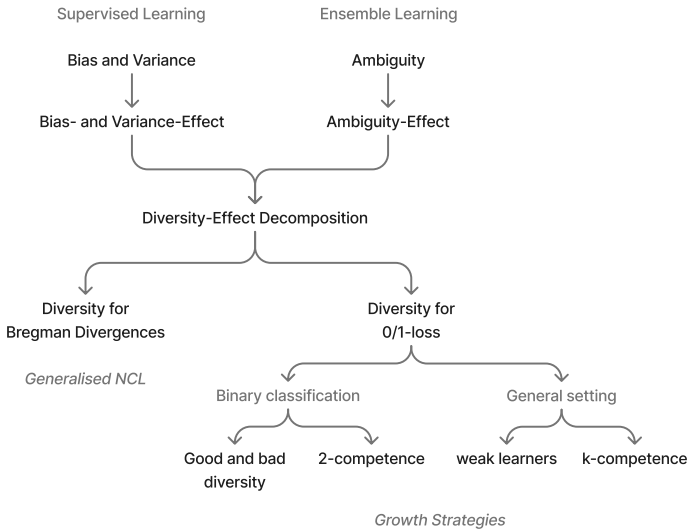
Model Correlation in Random Forests


Supervised Learning
Ensemble Learning
Bias and Variance
Ambiguity
Bias- and Variance-Effect
Ambiguity-Effect
Diversity-Effect Decomposition
Diversity for
Bregman Divergences
Diversity for
0/1-loss
Generalised NCL
Binary classification
General setting
Good and bad
diversity
2-competence
weak learners
k-competence
Growth Strategies

Supervised Learning objective

- Given features (description) of an object, predict an outcome for it.
- How to predict? Based on set of examples which we already know the outcome.
- Infer abstract *model* of unknown data from known training data

$\rightarrow$ Find a learning algorithm $\mathcal{A}$ that can be expected to produce "good" models.

Application example: Virtual Screening

- Many different cancer cell lines tested for their response to certain drugs
- Predict drug response for new cell lines

Determine promising candidates for more expensive *in-vitro* screening.

We need:

- Language and measures to assess learning algorithms

Setup : $D \sim P(\underbrace{\mathcal{X}}_{\text{Features}}, \underbrace{\mathcal{Y}}_{\text{Outcomes}})^n$

Training : $\underbrace{D}_{\text{Input}} \rightarrow \underbrace{\mathcal{A}}_{\text{Learner}} \rightarrow \underbrace{q_D}_{\text{Model}}$

Evaluation: $(X, Y) \sim P(\mathcal{X}, \mathcal{Y}); \quad \ell(q_D(X), Y)$

Loss / Error

Function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ to assess the error between two outcomes (e.g. true and predicted).

Model

Function $q : \mathcal{X} \rightarrow \mathcal{Y}$. In supervised learning, the model depends on the training input D. Model prediction:

$$q_D(X)$$

Risk and Generalisation Error

The *risk* of a model q_D is the expected loss over all example-outcome pairs.

$$\mathsf{Risk}(q_D) =_{\mathsf{def}} \mathbb{E}_{(X,Y)\sim P}\left[\ell(Y, q_D(X))\right]$$

The quality of a given learning algorithm $\mathcal{A}$ is the expected risk over all possible inputs D. We refer to this as the *generalisation error*.

$$\mathsf{GE}(\mathcal{A}) =_{\mathsf{def}} \mathbb{E}_D\left[\mathsf{Risk}(q_D)\right] = \mathbb{E}_{(X,Y),D}\left[\ell(Y, q_D(X))\right]$$

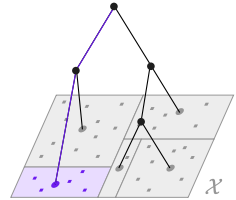
- Risk is property of model
- Generalisation error is property of learner

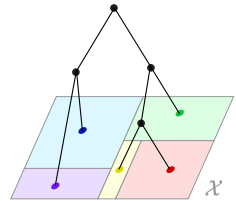
Basic idea: model prediction should be guided by outcomes of "similar" training examples.

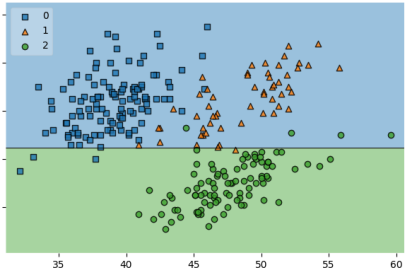
Decision Tree learner

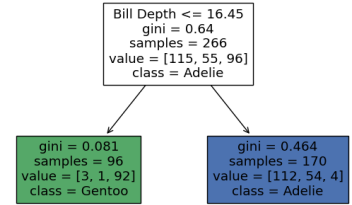
- Recursively perform binary splits in feature space $\mathcal{X}$
- Split such that "purity" of outcomes in cell is improved
- To predict, use centroid prediction of training examples in cell

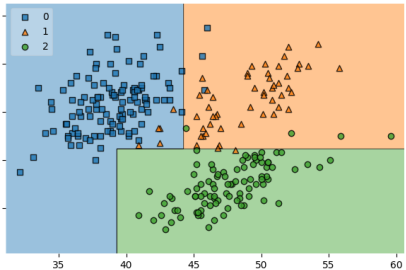
Prediction is cell-wise constant

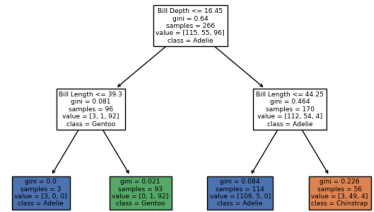


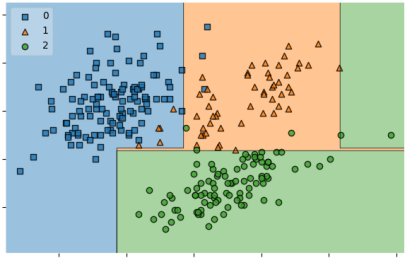


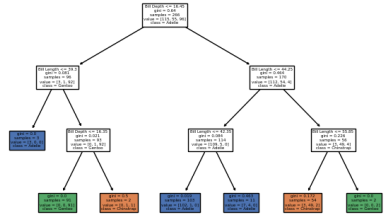

0
1
2
22
20
18
16
14
35
40
45
50
55
60

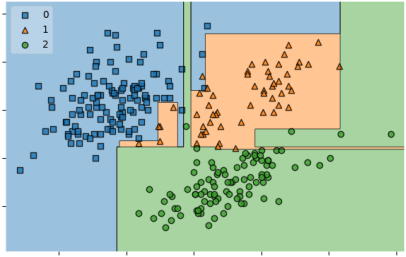

Bill Depth <= 16.45
gini = 0.64
samples = 266
value = [115, 55, 96]
class = Adelie
gini = 0.081
samples = 96
value = [3, 1, 92]
class = Gentoo
gini = 0.464
samples = 170
value = [112, 54, 4]
class = Adelie

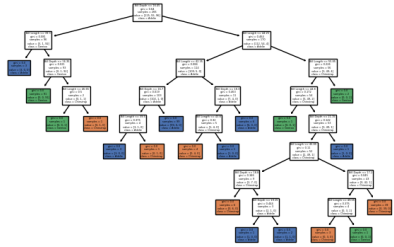

0
1
2
22
20
18
16
14
35
40
45
50
55
60


Bill Depth <= 16.45
gini = 0.64
samples = 266
value = [115, 55, 96]
class = Adelie
Bill Length <= 39.3
gini = 0.081
samples = 96
value = [3, 1, 92]
class = Gentoo
Bill Length <= 44.25
gini = 0.464
samples = 170
value = [112, 54, 4]
class = Adelie
gini = 0.0
samples = 3
value = [3, 0, 0]
class = Adelie
gini = 0.021
samples = 93
value = [0, 1, 92]
class = Gentoo
gini = 0.084
samples = 114
value = [109, 5, 0]
class = Adelie
gini = 0.226
samples = 56
value = [3, 49, 4]
class = Chinstrap

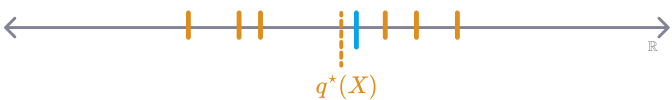

0
1
2
22
20
18
16
14
35
40
45
50
55
60




0
1
2
22
20
18
16
14
35
40
45
50
55
60



$$\mathsf{bias}(X, Y) =_{\mathsf{def}} \ell(Y, q^\star)$$



$$\begin{aligned}\mathsf{variance\text{-}effect}(X, Y) &=_{\mathsf{def}} \mathbb{E}_D\left[\ell(Y, q_D)\right] - \ell(Y, q^\star) \\ &= \mathbb{E}_D\left[\ell(Y, q_D) - \ell(Y, q^\star)\right] \\ &= \mathbb{E}_D\left[\mathsf{LE}(q^\star, q_D)\right]\end{aligned}$$

Central Model

$$q^\star =_{\mathsf{def}} \arg\min_z \mathbb{E}_D\left[\ell(z, q_D)\right]$$

Loss-Effect

For a loss funcion ℓ, and random variables Y, Z, Z', we define the change in loss between Z and Z' in relation to Y as:

$$\mathsf{LE}(Z, Z') =_{\mathsf{def}} \ell(Y, Z') - \ell(Y, Z)$$

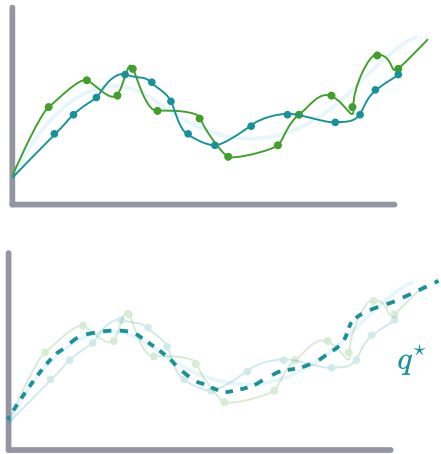
Bias-Variance-Effect-Decomposition for single point (X, Y) (no noise)

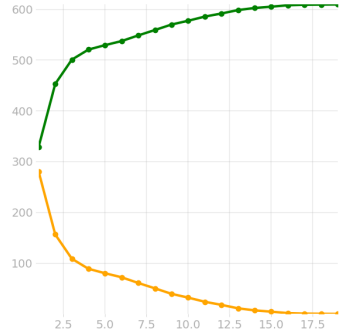
For any loss function ℓ, it holds that

$$\begin{aligned}\mathbb{E}_D\left[\ell(Y, q_D)\right] &= \ell(Y, q^\star) + \mathbb{E}_D\left[\ell(Y, q_D) - \ell(Y, q^\star)\right] \\ &= \underbrace{\ell(Y, q^\star)}_{\text{bias}} + \underbrace{\mathbb{E}_D\left[\mathsf{LE}(q^\star, q_D)\right]}_{\text{variance-effect}}\end{aligned}$$

For the squared-error loss $\ell(Z, Z') = (Z - Z')^2$, variance-effect = variance.

$$\mathbb{E}\left[\ell(Y, q) - \ell(Y, q^\star)\right] = \mathbb{E}\left[\ell(q^\star, q)\right]$$





Bias and variance of decision tree models of increasing maximum tree depth. With increasing tree depth, bias tends to decrease, as variance tends to increase.

Deep decision trees achieve low bias, but have high variance.

A Random Forest is an ensemble of M **randomised** decision trees.

Member learner (randomized with RV Θ, for e.g. single tree in forest)

$$\mathcal{A}_{\mathsf{DT}}(D, \Theta) \quad \rightsquigarrow \quad q_{D,\Theta} : \mathcal{X} \to \mathcal{Y}$$

Random Forest learner

$$\mathcal{A}_{\mathsf{RF}}(D) \quad \rightsquigarrow \quad \bar{q}_D : \mathcal{X} \to \mathcal{Y} : x \mapsto \arg\min_z \mathbb{E}_\Theta \left[\ell(z, q_{D,\Theta}(x))\right]$$

Implementation of $\mathcal{A}_{\mathsf{RF}}$:

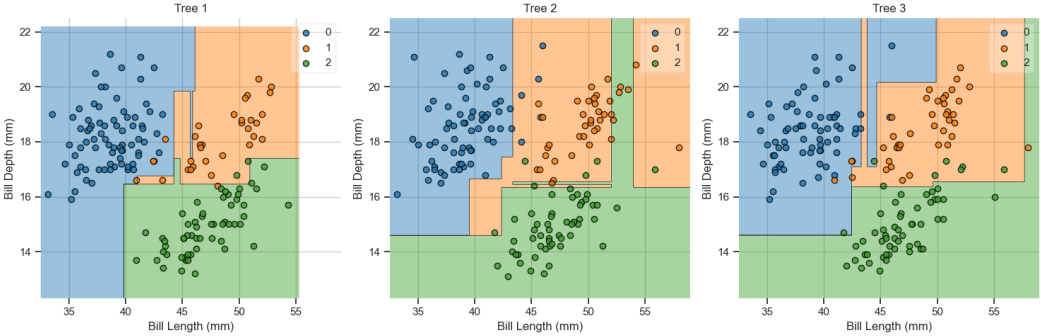
- Given D, repeat
  1. Draw Θ
  2. Construct a tree $q_{D,\Theta} \leftarrow \mathcal{A}_{\mathsf{DT}}(D, \Theta)$
- return $\bar{q}_D = \arg\min_z \mathbb{E}_\Theta \left[\ell(z, q_\Theta)\right]$

Randomness Θ used for

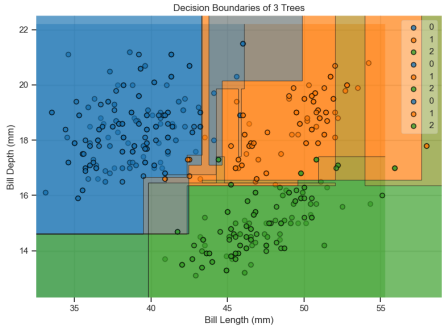
- Construct tree on random subset of D
- Only consider random subsample of feature dimensions for split search

$$D \sim P(\underbrace{\mathcal{X}}_{\text{Features}}, \underbrace{\mathcal{Y}}_{\text{Labels}})^n$$

$$\underbrace{D}_{\text{Input}} \to \underbrace{\mathcal{A}}_{\text{Learner}} \to \underbrace{q_D}_{\text{Model}}$$



Trees constructed using 3 different realisations of Θ. Scattered points are the bootstrap samples of D.



Ensemble Combiner

The ensemble combiner $\bar{q} : \mathcal{X} \to \mathcal{Y}$ for a given loss function ℓ is the centroid with respect to model parameters Θ:

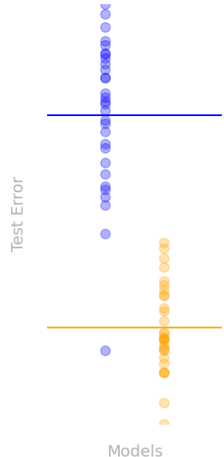
$$\bar{q} =_{\text{def}} \arg\min_{z} \mathbb{E}_{\Theta} \left[\ell(z, q_{\Theta}) \right]$$

- squared-error-loss:
  $\bar{q} = \mathbb{E}_{\Theta} \left[q_{\Theta} \right] \approx \frac{1}{M} \sum_{i=1}^{M} q_i$
- $0/1$-loss: $\bar{q}$ is plurality vote
- $\bar{q}$ is central model w.r.t. distribution of member parameters Θ

- The Random Forest learner has lower variance than the decision tree learner
- ... while bias is not affected "too much"

Questions

- When and why do such ensemble techniques work?
- In particular, what makes Random Forests work so well?



Ensemble Improvement / Diversity-Effect

The *ensemble improvement* is the difference in loss between the ensemble combiner and an average member.

$$\mathbb{E}_\Theta\left[\ell(Y, q_\Theta)\right] - \ell(Y, \bar{q})) = \mathbb{E}_\Theta\left[\mathsf{LE}(\bar{q}, q_\Theta)\right]$$

$$\text{variance-effect}(X, Y) =_{\text{def}} \mathbb{E}_D\left[\mathsf{LE}(q^\star, q_D)\right] \qquad q^\star =_{\text{def}} \arg\min_z \mathbb{E}_D\left[\ell(z, q_D)\right]$$

$$\text{diversity-effect}(X, Y) =_{\text{def}} \mathbb{E}_\Theta\left[\mathsf{LE}(\bar{q}, q_\Theta)\right] \qquad \bar{q} =_{\text{def}} \arg\min_z \mathbb{E}_\Theta\left[\ell(z, q_\Theta)\right]$$

Ambiguity-Effect decomposition

For any loss function ℓ, target label Y, ensemble members $q_1, \ldots, q_M$ with combiner $\bar{q}$

$$\begin{aligned}\ell(Y, \bar{q}) &= \mathbb{E}_\Theta\left[\ell(Y, q_\Theta)\right] - \mathbb{E}_\Theta\left[\ell(Y, q_\Theta) - \ell(Y, \bar{q})\right] \\ &= \mathbb{E}_\Theta\left[\ell(Y, q_\Theta)\right] - \mathbb{E}_\Theta\left[\mathsf{LE}(\bar{q}, q_\Theta)\right]\end{aligned}$$

Trade-off between average member error and diversity

Start with ambiguity decomposition

$$\mathbb{E}_D\left[\ell(Y,\bar{q})\right] = \mathbb{E}_{D,\Theta}\left[\ell(Y,q)\right] - \mathbb{E}_{D,\Theta}\left[\mathsf{LE}(\bar{q},q)\right]$$

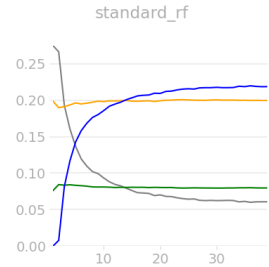
... and apply bias-variance decomp. to *member* loss $\ell(Y,q)$

Bias-Variance-Diversity-Effect decomposition

$$\mathbb{E}_D\left[\ell(y,\bar{q})\right] = \underbrace{\mathbb{E}_\Theta\left[\ell(Y,q^\star)\right]}_{\text{avg. bias}} + \underbrace{\mathbb{E}_{D,\Theta}\left[\mathsf{LE}(q^\star,q)\right]}_{\text{avg. variance-effect}} - \underbrace{\mathbb{E}_{D,\Theta}\left[\mathsf{LE}(\bar{q},q)\right]}_{\text{diversity-effect}}$$

- A model q depends on both D, Θ
- An expected/central model $q^\star = \arg\min_z \mathbb{E}_D\left[\ell(z,q)\right]$ depends on Θ only
- The combiner $\bar{q} = \arg\min_z \mathbb{E}_\Theta\left[\ell(z,q)\right]$ depends on D only

(diagram of "double decomp trick" like in wood23)

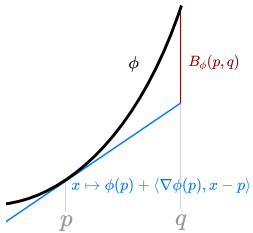


Bregman Divergence

The Bregman divergence $B_\phi(p,q) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is defined based on a generator function ϕ as follows:

$$B_\phi(p,q) =_{\text{def}} \phi(p) - \phi(q) - \langle \nabla\phi(q), (p-q) \rangle$$

where $\langle \cdot, \cdot \rangle$ is the inner product, $\nabla\phi(q)$ is the gradient vector of ϕ at q and $\phi : \mathcal{S} \to \mathbb{R}$ is a strictly convex function on a convex set $\mathcal{S} \subseteq \mathbb{R}^k$ such that it is differentiable on the relative interior of $\mathcal{S}$.



Divergence $B_\phi(p,q)$	Generator $\phi(q)$	Domain S	Loss function
$(p-q)^2$	q^2	$\mathbb{R}$	Squared Error
$p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$	$p\log p + (1-p)\log(1-p)$	$[0,1]$	Logistic loss
$\frac{p}{q} - \log\left(\frac{p}{q}\right) - 1$	$-\log p$	$\mathbb{R}_{>0}$	Ikura-Saito distance
$\|\| p-q \|\|^2$	$\|\| p \|\|^2$	$\mathbb{R}^d$	Squared Euclidean distance
$(p-q)^\top A(p-q)$	$p^\top Aq$	$\mathbb{R}^d$	Mahalanobis distance
$\sum_{j=1}^d p_j \log_2\left(\frac{p_j}{q_j}\right)$	$\sum_{j=1}^d p_j \log_2 p_j$	d-simplex	KL-divergence
$\sum_{j=1}^d p_j \log\left(\frac{p_j}{q_j}\right) - \sum_{j=1}^d (p_j - q_j)$	$\sum_{j=1}^d p_j \log p_j$	$\mathbb{R}^d_{\geq 0}$	Generalized I-divergence
$\sum_{j=1}^d p_j \log p_j$	$\sum_{j=1}^d p_j \log p_j$	$\mathbb{R}_{\geq 0}$	Poisson loss

Left and right Bregman centroids

Let B_ϕ be a Bregman divergence of generator $\phi : \mathcal{S} \to \mathbb{R}$. For a random variable Y taking values in $\mathcal{S}$, it holds that

- the *right Bregman centroid* is
$$\arg\min_z \mathbb{E}_X \left[B_\phi \left(X, z \right) \right] = \mathbb{E}\left[X \right]$$
- the *left Bregman centroid* is
$$\arg\min_z \mathbb{E}_X \left[B_\phi \left(z, X \right) \right] = (\nabla\phi)^{-1} \mathbb{E}\left[\nabla\phi(X) \right] =_{\mathsf{def}} \mathcal{E}\left[X \right]$$

If B_ϕ is symmetric, i.e. $B_\phi \left(Y, Y' \right) = B_\phi \left(Y', Y \right)$ then $\mathbb{E}\left[X \right] = \mathcal{E}\left[X \right]$.

Dual expectation

The left Bregman centroid is the expected value in the dual space implied by $\nabla\phi$. Due to this, we define the dual expectation as
$$\mathcal{E}\left[X \right] =_{\mathsf{def}} (\nabla\phi)^{-1} \mathbb{E}\left[\nabla\phi(X) \right]$$

For Bregman divergences, $q^\star$ and $\bar{q}$ are left Bregman centroids.

Let q be a function of random variable Z and independent of Y. For $q^\star = \mathcal{E}_Z[q]$, i.e. the left Bregman centroid w.r.t. Z, it holds that

$$\mathbb{E}_Z\left[B_\phi\left(q^\star, q\right)\right] = \mathbb{E}_{Z,Y}\left[B_\phi\left(Y, q\right) - B_\phi\left(Y, q^\star\right)\right]$$

For $q^\star = \mathcal{E}_D[q_D]$: $\mathbb{E}\left[B_\phi\left(q^\star, q\right)\right] = \mathbb{E}\left[\mathsf{LE}(q^\star, q)\right] = \mathbb{E}\left[B_\phi\left(Y, q\right) - B_\phi\left(Y, q^\star\right)\right]$

For $\bar{q} = \mathcal{E}_\Theta[q_\Theta]$: $\mathbb{E}\left[B_\phi\left(\bar{q}, q\right)\right] = \mathbb{E}\left[\mathsf{LE}(\bar{q}, q)\right] = \mathbb{E}\left[B_\phi\left(Y, q\right) - B_\phi\left(Y, \bar{q}\right)\right]$

Well-known bias-variance decomposition for squared-error loss is special case.

$$\mathbb{E}_D\left[\ell(y,\bar{q})\right] = \underbrace{\mathbb{E}_\Theta\left[\ell(Y,q^\star)\right]}_{\text{avg. bias}} + \underbrace{\mathbb{E}_{D,\Theta}\left[\mathsf{LE}(q^\star,q)\right]}_{\text{avg. variance-effect}} - \underbrace{\mathbb{E}_{D,\Theta}\left[\mathsf{LE}(\bar{q},q)\right]}_{\text{diversity-effect}}$$

Bias-Variance-Diversity decomposition for Bregman divergences

$$\mathbb{E}_D\left[\ell(y,\bar{q})\right] = \underbrace{\mathbb{E}_\Theta\left[\ell(Y,q^\star)\right]}_{\text{avg. bias}} + \underbrace{\mathbb{E}_{D,\Theta}\left[B_\phi\left(q^\star,q\right)\right]}_{\text{avg. variance}} - \underbrace{\mathbb{E}_{D,\Theta}\left[B_\phi\left(\bar{q},q\right)\right]}_{\text{diversity}}$$

Bregman divergences are non-negative $\rightarrow$ ensemble improvement / diversity-effect non-negative.

Under Bregman divergences, ...

- ... ensembling can not hurt performance (if using implied combiner)
- ... variance and diversity are independent of Y.

Homogeneous ensemble

An ensemble is homogeneous iff member parameters $\Theta_1, \ldots, \Theta_M$ are identically and independently distributed.

Let Θ, Θ' i.i.d. member parameters. In homogeneous ensembles the central models (w.r.t. D) of any two members and the combiner coincide

$$q^\star_\Theta = q^\star_{\Theta'} = \bar{q}^\star$$

if defined according to a Bregman divergence

(ensemble error) = (ensemble bias) + (ensemble-variance)

Unchanged bias and reduction in variance

- (ensemble bias) = (average member bias)
- (ensemble variance) = (average member variance) - (diversity)
- (average member variance) $\geq$ (diversity)

$0/1$-loss

$$\ell_{0/1}(Y, Y') =_{\text{def}} \mathbb{1}[Y \neq Y']$$

Majority/Plurality vote

$$\bar{q}(X) = \arg\min_{z \in [k]} \mathbb{E}_{\Theta}\left[\ell_{0/1}(z, q_{\Theta})\right]$$

Ratio of incorrect members

The expected ratio of incorrect ("wrong") members for a point (X, Y) is

$$W(X, Y) =_{\text{def}} \mathbb{E}_{D,\Theta}\left[\ell_{0/1}(Y, q_{D,\Theta}(X))\right]$$

We write $W_{\Theta} =_{\text{def}} \mathbb{E}_{\Theta}\left[\ell_{0/1}(Y, q_{D,\Theta})\right]$. For the complement, write $\overline{W} =_{\text{def}} 1 - W$

A simple but tight bound using Markov's inequality

$$0 \leq \mathbb{E}\left[\ell_{0/1}(Y, \bar{q})\right] \leq \mathbb{P}[W \geq 1/2] \leq 2\mathbb{E}[W]$$

$\rightsquigarrow$ need assumptions on performance of members

- For $0/1$-loss, $\mathsf{LE}(\bar{q}, q)$ not necessarily non-negative

$\rightarrow$ can have points (X, Y) with non-negative diversity-effect

$\rightarrow$ *hurts* ensemble generalisation error

$\rightarrow$ ensemble improvement not necessarily non-negative

Weak learner condition

A model q_Θ is a weak learner if and only if performs better than randomly guessing.

$$\mathbb{E}_{(X,Y)}\left[\ell_{0/1}\left(Y, q_\Theta\right)\right] \geq 1/2$$

In an ensemble of weak learners, diversity-effect/ensemble improvement is non-negative:

$$\mathbb{E}_{(X,Y),D,\Theta}\left[\ell_{0/1}\left(Y, q_i\right) - \ell_{0/1}\left(Y, \bar{q}\right)\right] \quad \geq \quad 0$$

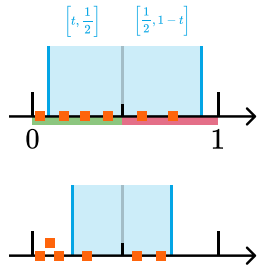
2-competence,

An ensemble is 2-competent iff

$$\forall t \in [0, 1/2]: \ \mathbb{P}_{(X,Y)}[W \in [t, 1/2]] \quad \geq \quad \mathbb{P}_{(X,Y)}[W \in [1/2, 1-t]]$$

Original results: In 2-competent ensembles

- 2-competence $\rightarrow$ (diversity-effect) ≥ 0
- ensemble error is bounded by linear functions of expected disagreement



Weak learner: $\mathbb{E}_{(X,Y)}\left[\ell_{0/1}\left(Y, q_\Theta\right)\right] \geq 1/2$

2-competence: $\forall t \in [0, 1/2]: \ \mathbb{P}_{(X,Y)}\left[W \in [t, 1/2]\right] \quad \geq \quad \mathbb{P}_{(X,Y)}\left[W \in [1/2, 1-t]\right]$

$\star$

The weak learner condition is a special case of 2-competence

We have also in expectation

$$\mathbb{E}_{\Theta,D}\left[\mathbb{E}_{(X,Y)}\left[\ell_{0/1}\left(Y, \bar{q}\right)\right]\right] \geq 1/2$$

Consequently,

$$\begin{aligned} 1/2 \leq \mathbb{E}_{\Theta,D,(X,Y)}\left[\ell_{0/1}\left(Y, \bar{q}\right)\right] &= \mathbb{E}_{(X,Y)}\left[W\right] \\ \leftrightarrow \ \mathbb{E}_{(X,Y)}\left[\ \mathbb{1}\left[W \in [0, 1/2]\right]\right] &= \mathbb{P}_{(X,Y)}\left[W \in [0, 1/2]\right] = 1 \end{aligned}$$

Assume $k = 2$. Any vote that is not for class y is automatically for the single other class.

$$k = 2 \;\rightarrow\; \begin{cases} W_\Theta < 1/2 \;\leftrightarrow\; \bar{q}(X) = Y \\ \overline{W_\Theta} \leq 1/2 \;\leftrightarrow\; \bar{q}(X) \neq Y \end{cases}$$

Starting from ambiguity decomp.:

Lemma

$$\begin{aligned} \mathbb{E}_{(X,Y)}\left[\ell(Y,\bar{q})\right] &= \mathbb{E}_{(X,Y),\Theta}\left[\ell(Y,q)\right] - \mathbb{E}_{(X,Y),\Theta}\left[\ell(Y,q) - \ell(Y,\bar{q})\right] \\ &= \mathbb{E}_{(X,Y),\Theta}\left[\ell(Y,q)\right] - \left(\mathbb{E}_{X_+}\left[W_\Theta\right] + \mathbb{E}_{X_-}\left[W_\Theta - 1\right]\right) \\ &= \mathbb{E}_{(X,Y),\Theta}\left[\ell(Y,q)\right] \underbrace{- \mathbb{E}_{X_+}\left[W_\Theta\right] + \mathbb{E}_{X_-}\left[1 - W_\Theta\right]}_{\text{diversity-effect for } k=2} \end{aligned}$$

where X_+ is range of (X,Y) where $\bar{q}(X) = Y$ and X_- vice versa. Assume $Y = \bar{q}$, then $\ell_{0/1}(Y,\bar{q}) = 0$ and diversity-effect becomes W. Assume $Y \neq \bar{q}$, then diversity-effect becomes $\overline{W} - 1 = 1 - W$

Theisen et al. have shown that (2-competence) $\rightarrow$ (diversity-effect) ≥ 0. They established:

$$\text{2-competence} \;\leftrightarrow\; \mathbb{E}\left[W\ \mathbb{1}\left[W < 1/2\right]\right] \geq \mathbb{E}\left[\overline{W}\ \mathbb{1}\left[\overline{W} \leq 1/2\right]\right]$$

Consequently

$$d =_{\text{def}} \mathbb{E}_{(X,Y),D}\left[W_\Theta\ \mathbb{1}\left[W_\Theta < 1/2\right]\right] - \mathbb{E}_{(X,Y),D}\left[\overline{W_\Theta}\ \mathbb{1}\left[\overline{W_\Theta} \leq 1/2\right]\right] \quad \geq \quad 0$$

The indicator functions are mutually exclusive and can be understood as a case distinction.

$$d = \begin{cases} \mathbb{E}_{X_+}\left[W_\Theta\right] & \leftrightarrow W_\Theta < 1/2 \\ \mathbb{E}_{X_-}\left[\overline{W_\Theta}\right] = \mathbb{E}\left[1 - W_\Theta\right] & \leftrightarrow \overline{W_\Theta} \leq 1/2 \end{cases}$$

In binary classification, d is nothing but diversity-effect.

$\star$ In binary classification problems, 2-competence and non-negative diversity-effect are equivalent.

$$k = 2 \quad \Rightarrow \quad (\text{2-competence} \leftrightarrow \text{diversity-effect} \geq 0)$$

- For $k = 2$, $1/2$ is the classification threshold.
- For $k > 2$, $W \leq 1/2$ is sufficient but not necessary for correctness.
- Plurality vote can be won with less than $1/2M$ votes.

Effective classification threshold for $k > 2$ is number of votes for next-best class.

$$\kappa(X, Y) = 1 - \max_{Z \neq Y} \mathbb{E}_\Theta \left[\mathbb{1} \left[q_\Theta = Z \right] \right]$$

$$k \text{ arbitrary } \rightarrow \begin{cases} W_\Theta < \kappa & \leftrightarrow \ \bar{q}(X) = Y \\ \overline{W_\Theta} \leq 1 - \kappa & \leftrightarrow \ \bar{q}(X) \neq Y \end{cases}$$

Claim

We can work with just *some* classification threshold

Recap: An ensemble is 2-competent iff

$$\forall t \in [0, 1/2] : \ \mathbb{P}_{(X,Y)}\left[W \in [t, 1/2]\right] \quad \geq \quad \mathbb{P}_{(X,Y)}\left[W \in [1/2, 1-t]\right]$$

k-competence

An ensemble is k-competent iff

$$\forall t \in [0, 1] : \mathbb{P}_{(X,Y)}\left[W \in [t, \kappa]\right] \geq \mathbb{P}_{(X,Y)}\left[W \in [1-\kappa, 1-t]\right]$$

for $\kappa =_{\text{def}} 1 - \max_{Z \neq Y} \mathbb{E}_{\Theta}\left[\ \mathbb{1}\left[q_{\Theta} = Z\right]\right]$.

$\star$ 2-competence is a special case of k-competence in 2-class problems.

We have seen that

$$k = 2: \quad \text{2-competence} \ \leftrightarrow \ \text{diversity-effect} \geq 0$$

and we will show shortly that

$$\text{diversity-effect} \geq 0 \ \leftrightarrow \ k\text{-competence}$$

Recap:

$$2\text{-competence} \;\leftrightarrow\; \mathbb{E}\left[W\ \mathbb{1}\left[W < 1/2\right]\right] \geq \mathbb{E}\left[\overline{W}\ \mathbb{1}\left[\overline{W} \leq 1/2\right]\right]$$

Lemma

$\star$ It holds that

$$k\text{-competence} \;\leftrightarrow\; \mathbb{E}\left[W\ \mathbb{1}\left[W < \kappa\right]\right] \geq \mathbb{E}\left[\overline{W}\ \mathbb{1}\left[\overline{W} \leq \kappa\right]\right]$$

where $\kappa =_{\text{def}} 1 - \max_{Z \neq Y} \mathbb{E}_{\Theta}\left[\ \mathbb{1}\left[q_{\Theta} = Z\right]\right]$.

We begin by observing that, for all $x \in [0,1]$

$$\mathbb{P}\left[W \in [x,\kappa]\right] \cdot\ \mathbb{1}\left[x \leq \kappa\right] = \mathbb{P}\left[W\ \mathbb{1}\left[W < \kappa\right]\ \ \geq\ \ x\right]$$
$$\mathbb{P}\left[W \in [1-\kappa, 1-x]\right] \cdot\ \mathbb{1}\left[x \leq \kappa\right] = \mathbb{P}\left[\overline{W}\ \mathbb{1}\left[\overline{W} \leq \kappa\right]\ \ \geq\ \ x\right]$$

where the first factors on the left-hand-side appear in the definition of k-competence. Since W is nonnegative, using that $\mathbb{E}\left[X\right] = \int \mathbb{P}\left[X \geq x\right] dx$, we can conclude that, for any $x \in [0,1]$

$$\begin{aligned}(k\text{-comp.}) \;&\leftrightarrow\; \mathbb{P}\left[W\ \mathbb{1}\left[W < \kappa\right]\ \ \geq x\right] \quad\geq\quad \mathbb{P}\left[\overline{W}\ \mathbb{1}\left[\overline{W} \leq \kappa\right]\ \ \geq x\right] \\ &\leftrightarrow\; \mathbb{E}\left[W\ \mathbb{1}\left[W < \kappa\right]\right] \quad\geq\quad \mathbb{E}\left[\overline{W}\ \mathbb{1}\left[\overline{W} \leq \kappa\right]\right]\end{aligned}$$

⋆ Consider an ensemble for a k-class classification problem. Then

$$k\text{-competence} \quad \leftrightarrow \quad \text{diversity-effect} \geq 0$$

$$\begin{aligned}
0 &= \mathbb{E}\left[(W-1)\ \mathbb{1}\left[W \geq \kappa\right]\right] - \mathbb{E}\left[(W-1)\ \mathbb{1}\left[W \geq \kappa\right]\right] \\
&= \mathbb{E}\left[(W-1)\ \mathbb{1}\left[W \geq \kappa\right]\right] + \mathbb{E}\left[(1-W)\ \mathbb{1}\left[W \geq \kappa\right]\right] \\
&= \mathbb{E}\left[(W-1)\ \mathbb{1}\left[W \geq \kappa\right]\right] + \mathbb{E}\left[\overline{W}\ \mathbb{1}\left[\overline{W} < 1-\kappa\right]\right] \\
&\leq \mathbb{E}\left[(W-1)\ \mathbb{1}\left[W \geq \kappa\right]\right] + \mathbb{E}\left[\overline{W}\ \mathbb{1}\left[\overline{W} < \kappa\right]\right]
\end{aligned}$$

Where the final inequality is enabled due to that, for $k \geq 2$, we have $\max_{Z \neq Y} \mathbb{E}_{\Theta}\left[\ \mathbb{1}\left[q_{\Theta} = Z\right]\right] < 1/2$ and consequently $\kappa \geq 1/2$ and $\kappa > 1-\kappa$. Applying the lemma to the second term yields

$$\begin{aligned}
&\mathbb{E}\left[(W-1)\ \mathbb{1}\left[W \geq \kappa\right]\right] + \mathbb{E}\left[\overline{W}\ \mathbb{1}\left[\overline{W} < \kappa\right]\right] \\
\leq\ &\mathbb{E}\left[(W-1)\ \mathbb{1}\left[W \geq \kappa\right]\right] + \mathbb{E}\left[W\ \mathbb{1}\left[W < \kappa\right]\right]
\end{aligned}$$

The above already is nothing but the diversity-effect:

$$\begin{aligned} 0 \leq & \mathbb{E}\left[(W-1)\ \mathbb{1}\left[W \geq \kappa\right]\right] + \mathbb{E}\left[W\ \mathbb{1}\left[W < \kappa\right]\right] \\ = & \mathbb{E}\left[W\right] - \mathbb{E}\left[\ \mathbb{1}\left[W \geq \kappa\right]\right] \\ = & \mathbb{E}\left[W\right] - \mathbb{P}\left[W \geq \kappa\right] \end{aligned}$$

The first term is the ratio of incorrect members in expectation over all examples and is equal to the error rate of an average member. The second term is the ensemble error.

Note

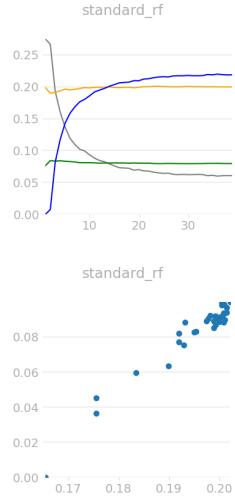
Diversity is a measure of model fit (just like bias and variance)

$\rightsquigarrow$ A more diverse ensemble is not necessarily better
Nevertheless:

Questions

- Can we encourage diversity in Random Forests?
- Can we produce Random Forests with better generalisation error?
- Can we produce smaller forests?

We consider $0/1$-classification.



Random Forests with varying number of trees plotted across average member error (vertical axis) and diversity (horizontal axis).

Recall ambiguity-effect decomposition:

$$\mathbb{E}\left[\ell_{0/1}\left(Y,\bar{q}\right)\right] = \mathbb{E}\left[\ell_{0/1}\left(Y,q\right)\right] - \mathbb{E}\left[\ell_{0/1}\left(Y,q\right) - \ell_{0/1}\left(Y,\bar{q}\right)\right]$$

For some point (X, Y):

- Diversity term is positive (beneficial) if ensemble is correct
- Diversity term is negative (detrimental) if ensemble is incorrect

Basic idea: Encourage diversity on points where we can *afford* diversity. (Increase "competence gap")

Simple approach:

- Example weights $w(X) \rightsquigarrow$ will influence split search in DT construction
- Lower weight for $X \rightsquigarrow$ higher average member error on $X \rightsquigarrow$ higher diversity

- Decision tree construction (implicitly) optimises a global loss function ℓ_{DT}.
- Example weights correspond to "gradient descent in function space" ($\rightsquigarrow$ boosting).

Objective becomes

$$\ell_{DT}(q(X), Y) - \lambda \cdot \ell_{\mathsf{reg}}(q(X), Y)$$

DRF weighting scheme

Let $\bar{q}$ be the ensemble constructed so far. For a pair $(X, Y) \in D$, define the *Dynamic Random Forest weighting scheme* as

$$w_{\mathsf{DRF}}(X) =_{\mathsf{def}} W(X) \qquad \text{for } W(X) =_{\mathsf{def}} \frac{1}{M} \sum_{i=1}^{M} \ell_{0/1}\left(Y, q_i(X)\right)$$

$$w_{\mathsf{XuChen}}(X) =_{\mathsf{def}} \begin{cases} W(X)^2 & \text{if } W(X) \leq 1/2 \\ \sqrt{W(X)} & \text{if } W(X) > 1/2 \end{cases}$$

Note: only makes sense for binary classification (threshold at $1/2$)

Two options to use weights:

- Weighted bootstrapping
- Weighted tree construction

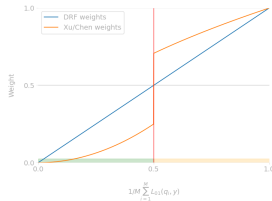
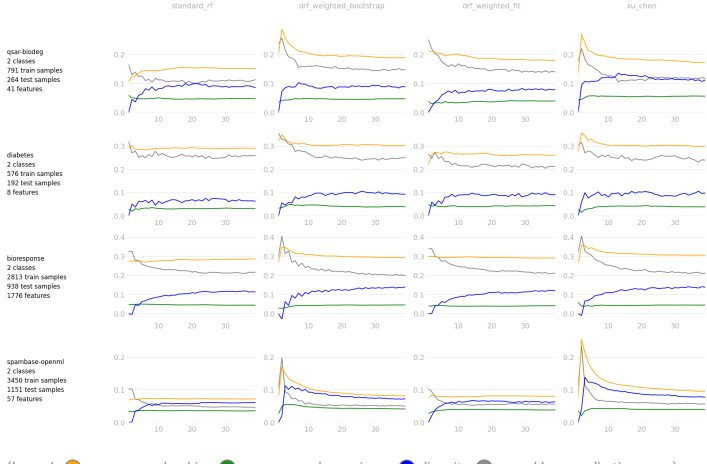


Illustration of w_{DRF} and w_{XuChen}.



- w_{DRF} and w_{XuChen} with weighted bootstrapping indeed lead to more diverse ensembles. For weighted tree construction, diversity is very similar to standard Random Forests.
- For weighted bootstrapping, there is an initial sharp increase in diversity and average member error.

Claim

- Incorrectly classified examples should not receive a higher weight
  - If anything, lower: subsequent correct votes will increase "bad" diversity (until majority vote is tipped)
- Weighting function does not have to be linear.

Sigmoid weighting function

$$
\begin{aligned}
w_{\text{sigm}}(X) &=_{\text{def}} \min\{s, 1/2\} \\
\text{for} \quad s &=_{\text{def}} \text{sigmoid}(W(X) - t; 0, b, 1) \\
t &=_{\text{def}} 1/2 \\
b &=_{\text{def}} b_{\text{max}}
\end{aligned}
$$

where the minimum clips the function values to a maximum of $1/2$, $t = 1/2$ is the voting threshold.

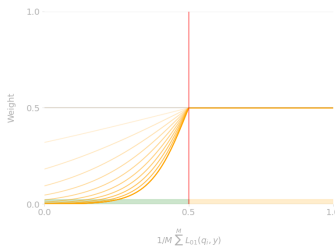
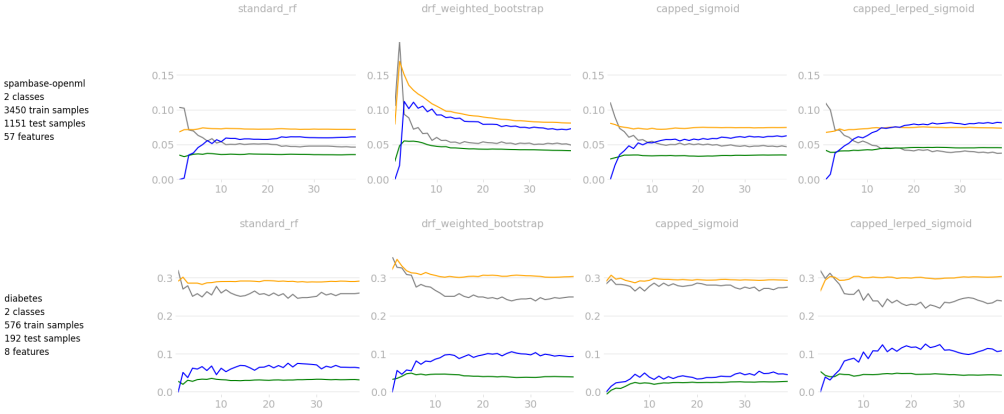


Illustration of w_{sigm} for $t = 1/2$ and varying b.

w_{lerp}

$$
b \leftarrow \frac{\min\{M, M_{\text{max}}\}}{M_{\text{max}}} b_{\text{max}}
$$



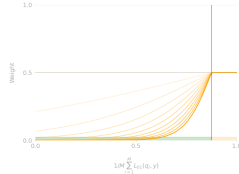
Comparison of w_{DRF}, w_{sigm} and w_{lerp}. *diabetes* is a very small dataset with a high best achievable error rate.

- Clipping (not assigning higher weights to incorrectly classified examples) already mitigates the spike in member error and diversity. Interpolating the steepness of the weighting function b is not required for this.
- Surprisingly, interpolating leads to higher diversity (sometimes even higher than w_{DRF}) and lower or similar ensemble generalisation error.

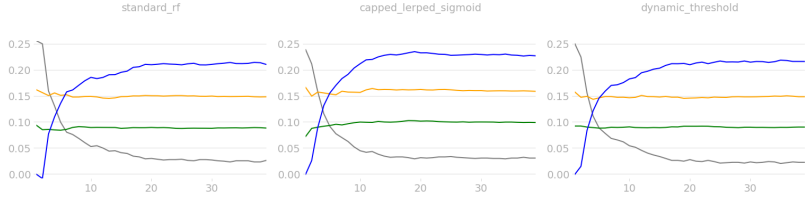
For non-binary classification problems:

w_{dyn} is defined analogously to w_{lerp}, but t is given as

$$t =_{\text{def}} \kappa(X, Y) = 1 - \max_{Z \neq Y} \mathbb{E}_{\Theta}[\, \mathbb{1}[q_{\Theta} = Z]]$$



digits
10 classes
1347 train samples
450 test samples
64 features

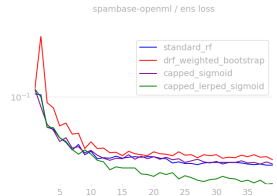


Comparison of w_{lerp} and w_{dyn} for a non-binary classification problem.

- Using w_{dyn}, one can achieve slighly improvided ensemble generalisation error as compared to any other learner.
- The development of diversity is slower than with w_{lerp} and more similar to standard Random Forests.

Generalisation error for binary datasets

	RF	w_{DRF}	w_{sigm}	w_{lerp}
qsar-biodeg	**0.101**	0.144	0.130	0.133
diabetes	0.247	0.240	0.266	**0.220**
bioresponse	0.211	**0.198**	0.217	0.209
spambase-openml	0.047	0.050	0.047	**0.038**
digits	**0.024**	0.033	0.029	0.030
mnist-subset	**0.060**	0.066	0.064	0.062
cover	0.049	**0.042**	0.054	0.042



Generalisation error for non-binary datasets

	RF	w_{DRF}	w_{lerp}	w_{dyn}
digits	0.024	0.033	0.030	**0.021**
mnist-subset	0.060	0.066	0.062	**0.056**
cover	0.049	**0.042**	0.042	0.050

- ⋆ Bias-Variance-Diversity decomposition from first principles
- Bregman divergences vs. general loss functions and $0/1$-loss (non-negative vs good/bad diversity)
- Weak learners and competence
  - Weak learners $\rightarrow$ 2-competence
    - ⋆ $k = 2 \Rightarrow$ 2-competence $\leftarrow$ diversity-effect ≥ 0 (equivalent)
    - ⋆ $k = 2 \Rightarrow$ 2-competence $\rightarrow$ k-competence
    - ⋆ k arbitrary $\Rightarrow$ k-competence $\leftrightarrow$ diversity-effect ≥ 0
- Growth strategies
  - Possible to encourage diversity in $0/1$-classification RFs
  - Behaviour depends on datasets (error tradeoff; noise?)
  - Possible to obtain better, smaller ensembles than standard RF

⋆ Bias-Variance-Diversity decomposition from first principles

- Bregman divergences vs. general loss functions and $0/1$-loss (non-negative vs good/bad diversity)
- Weak learners and competence
  - Weak learners $\rightarrow$ 2-competence
    - ⋆ $k = 2 \Rightarrow$ 2-competence $\leftarrow$ diversity-effect ≥ 0 (equivalent)
    - ⋆ $k = 2 \Rightarrow$ 2-competence $\rightarrow$ k-competence
    - ⋆ k arbitrary $\Rightarrow$ k-competence $\leftrightarrow$ diversity-effect ≥ 0
- Growth strategies
  - Possible to encourage diversity in $0/1$-classification RFs
  - Behaviour depends on datasets (error tradeoff; noise?)
  - Possible to obtain better, smaller ensembles than standard RF

Thanks for listening

$\mathbb{E}_{(X,Y)}\left[\ell(Y, y^\star)\right]$ is *noise*.

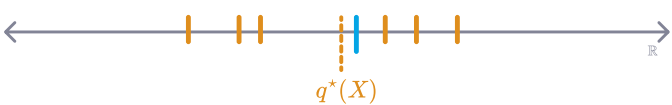
Central Label

$$y^\star(X) =_{\text{def}} \mathbb{E}_{Y|X}\left[Y\right]$$

$\text{bias-effect}(X, Y) =_{\text{def}} \ell(Y, q^\star) - \ell(Y, y^\star)$

Central Model

$$q^\star =_{\text{def}} \arg\min_z \mathbb{E}_D\left[\ell(z, q_D)\right]$$



$$\mathbb{E}_D\left[\ell(Y, q_D)\right] - \ell(Y, q^\star)$$
$$= \mathbb{E}_D\left[\ell(Y, q_D) - \ell(Y, q^\star)\right]$$

$\text{variance-effect}(X, Y) =_{\text{def}} \mathbb{E}_D\left[\ell(Y, q_D)\right] - \ell(Y, q^\star)$

$$\mathbb{E}\left[\ell(Y, q)\right] = \mathbb{E}\left[\ell(Y, y^\star) + \ell(Y, q^\star) - \ell(Y, y^\star) + \ell(Y, q) - \ell(Y, q^\star)\right]$$