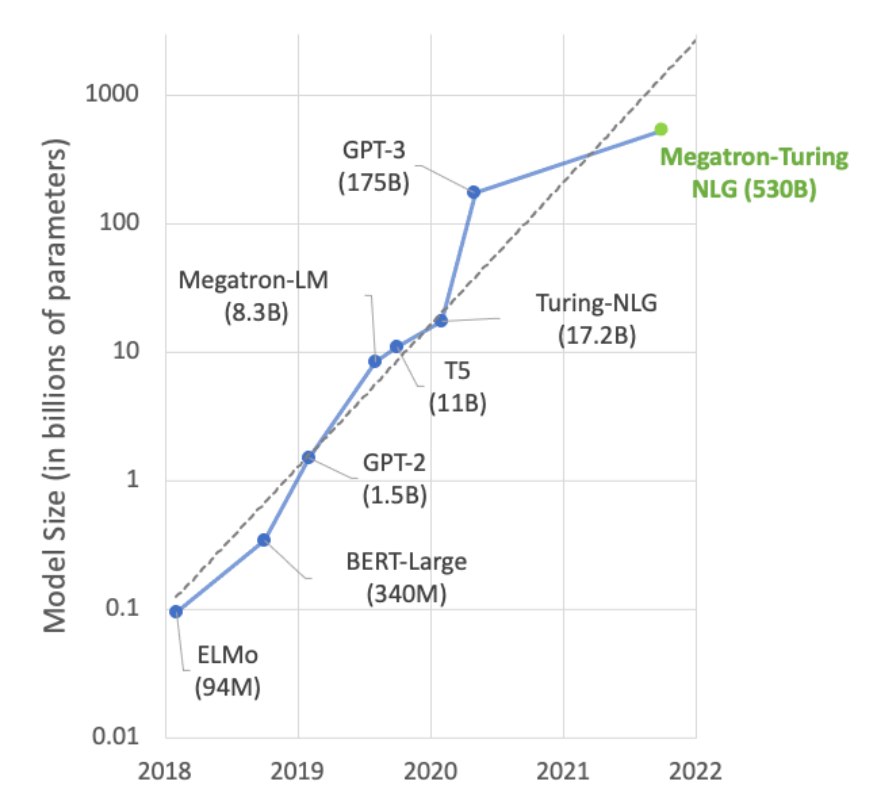
MATHEMATICAL TOPICS IN MACHINE LEARNING

(LECTURE 6 – ENSEMBLE LEARNING THEORY)

Professor Gavin Brown

OUR QUESTION

"Are bigger models always better models?"



TODAY:

We examine how bias & variance manifest when we use model ensembles.

TODAY…

Sections 1, 2, 3, and 4.1 are your assigned reading.

I will cover most,
but not all,
in this session.

https://jmlr.org/papers/volume24/23-0041/23-0041.pdf

Journal of Machine Learning Research 24 (2023) 1-49 Submitted 1/23; Revised 12/23; Published 12/23

A Unified Theory of Diversity in Ensemble Learning

Danny Wood DANNY.WOOD@MANCHESTER.AC.UK†
Tingting Mu TINGTING.MU@MANCHESTER.AC.UK†
Andrew M. Webb ANDREW.WEBB@MANCHESTER.AC.UK†
Henry W. J. Reeve HENRY.REEVE@BRISTOL.AC.UK*
Mikel Luján MIKEL.LUJAN@MANCHESTER.AC.UK†
Gavin Brown GAVIN.BROWN@MANCHESTER.AC.UK†
† *Department of Computer Science, University of Manchester, UK*
∗ *School of Mathematics, University of Bristol, UK*

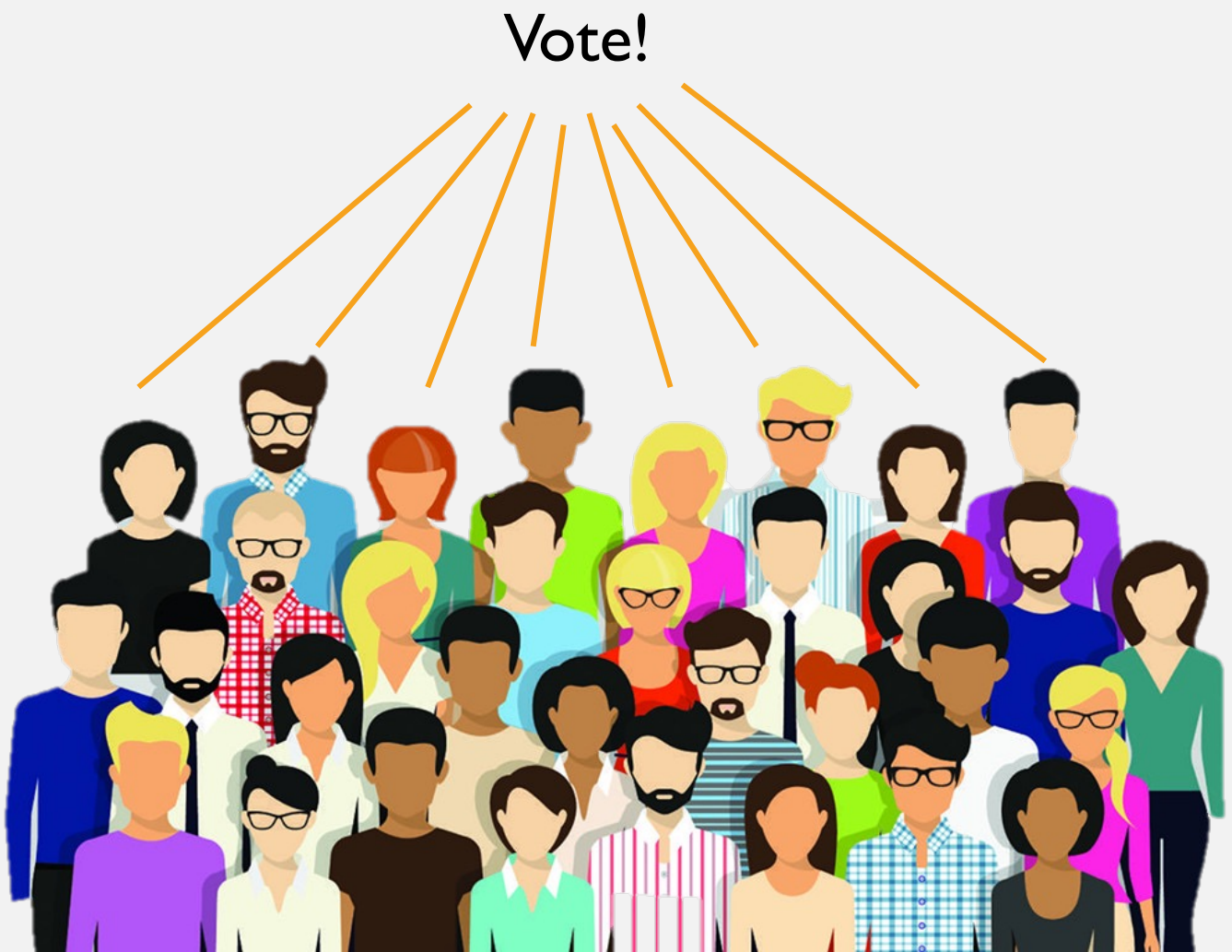
Editor: Boaz Nadler

Abstract

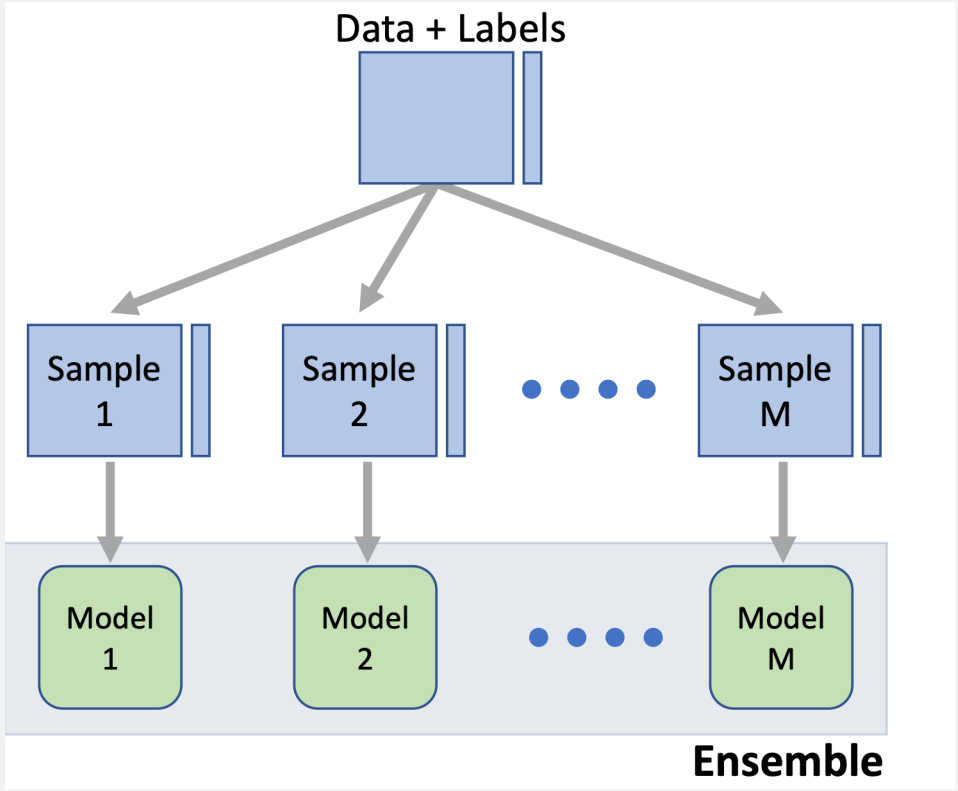
We present a theory of ensemble diversity, explaining the nature of diversity for a wide range of supervised learning scenarios. This challenge has been referred to as the "holy grail" of ensemble learning, an open research issue for over 30 years. Our framework reveals that diversity is in fact a *hidden dimension* in the bias-variance decomposition of the ensemble loss. We prove a family of *exact* bias-variance-diversity decompositions, for a wide range of losses in both regression and classification, e.g., squared, cross-entropy, and Poisson losses. For losses where an additive bias-variance decomposition is not available (e.g., 0/1 loss) we present an alternative approach: quantifying the *effects* of diversity, which turn out to

DIVERSE COMMITTEES ARE BETTER

We believe this. It's an unquestioned belief in modern society.

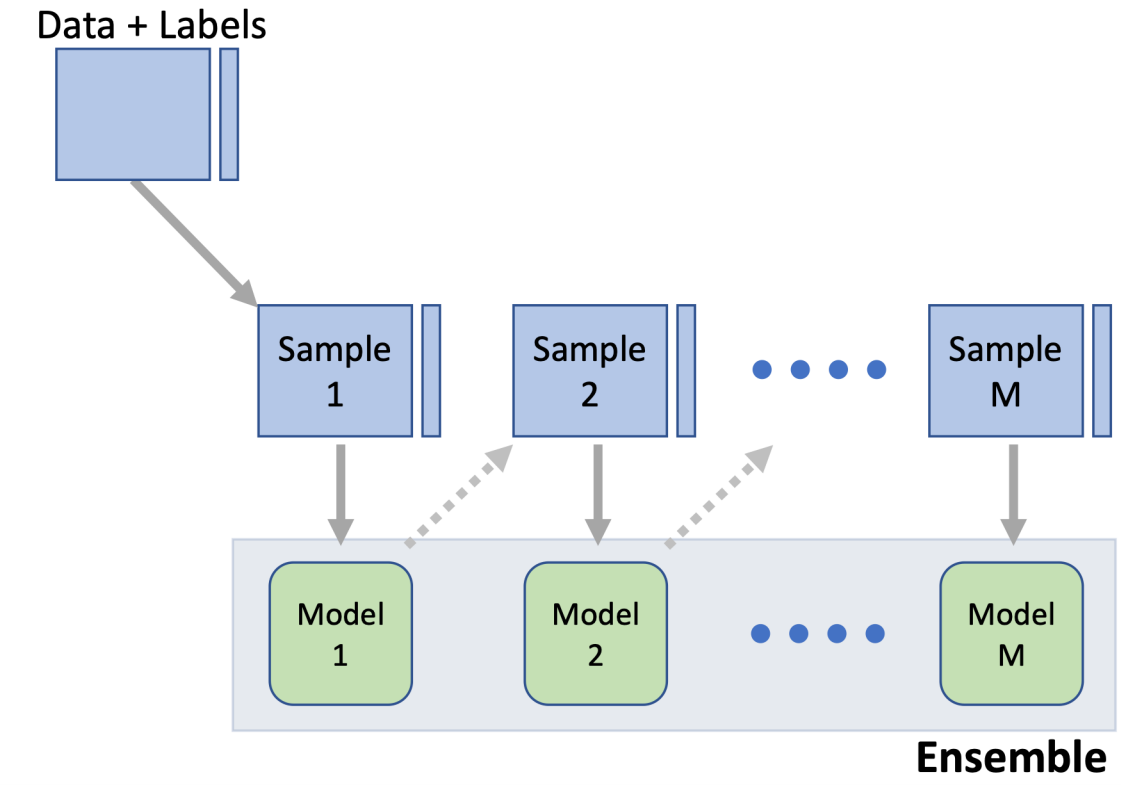


WE ADOPTED THIS **IDEA** IN MACHINE LEARNING



Bagging: encourages "diverse" predictions by providing different training samples to each model

WE ADOPTED THIS **IDEA** IN MACHINE LEARNING



Boosting: encourages "diverse" predictions by forcing each model to be accurate where earlier ones were not

A PARTIAL ANSWER FROM LAST WEEK

Squared loss

$$\Big(\bar{f}(\mathbf{x}) - y\Big)^2 = \frac{1}{m}\sum_{i=1}^{m}\Big(f_i(\mathbf{x}) - y\Big)^2 - \frac{1}{m}\sum_{i=1}^{m}\Big(f_i(\mathbf{x}) - \bar{f}(\mathbf{x})\Big)^2$$

Arithmetic mean

$$\bar{f}(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x})$$

Cross-entropy loss

$$\ell(\mathbf{y}, \bar{f}(\mathbf{x})) = \frac{1}{m}\sum_{i=1}^{m}\ell(\mathbf{y}, f_i(\mathbf{x})) - \frac{1}{m}\sum_{i=1}^{m} K(\bar{f}(\mathbf{x}), f_i(\mathbf{x}))$$

Geometric mean

$$\bar{f}(\mathbf{x}) = Z^{-1}\prod_{i} f_i(\mathbf{x})^{\frac{1}{m}}$$

Loss of the ensemble is guaranteed to be **less than or equal to** the average loss.
The reduction in loss is given by the "**ambiguity**" term.

REMINDER: BIAS/VARIANCE

The "expected" risk :

$$\mathbb{E}_{\mathcal{S}_n}\Big[R(f)\Big] = \mathbb{E}_{\mathcal{S}_n}\Big[\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\Big[(f(\boldsymbol{x})-y)^2\Big]\Big]$$

Squared loss of a **single** trained model,
at a **single** test point,
averaged over **all possible training sets**,
averaged over **all possible** test points.

REMINDER: BIAS/VARIANCE

$$\underbrace{\mathbb{E}_{\mathcal{S}_n}\Big[R(f)\Big]}_{\text{expected squared risk}} = \mathbb{E}_{\boldsymbol{x}}\Bigg[\underbrace{\mathbb{E}_{y|\boldsymbol{x}}\Big[\big(y-\mathbb{E}_{y|\boldsymbol{x}}[y]\big)^2\Big]}_{\text{noise}} + \underbrace{\Big(\mathbb{E}_{\mathcal{S}_n}\left[f(\boldsymbol{x})\right]-\mathbb{E}_{y|\boldsymbol{x}}[y]\Big)^2}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathcal{S}_n}\Big[\Big(f(\boldsymbol{x})-\mathbb{E}_{\mathcal{S}_n}\left[f(\boldsymbol{x})\right]\Big)^2\Big]}_{\text{variance}}\Bigg].$$

The noise is a constant that we cannot control. It is the population risk of the Bayes model, i.e., $R(y^*)$.

The squared difference between the "expected" model and the Bayes model.

The variation of the trained model, the expected model, in response to getting different training sets.

The "sensitivity" to its training data.

Notice that the model is not mentioned in this analysis.

$\mathbb{E}_{\mathcal{S}_n}\left[f(\mathbf{x})\right]$

"**Expected**" model = the response that the model would give if we could average over all possible training sets.

THE SHORT STORY

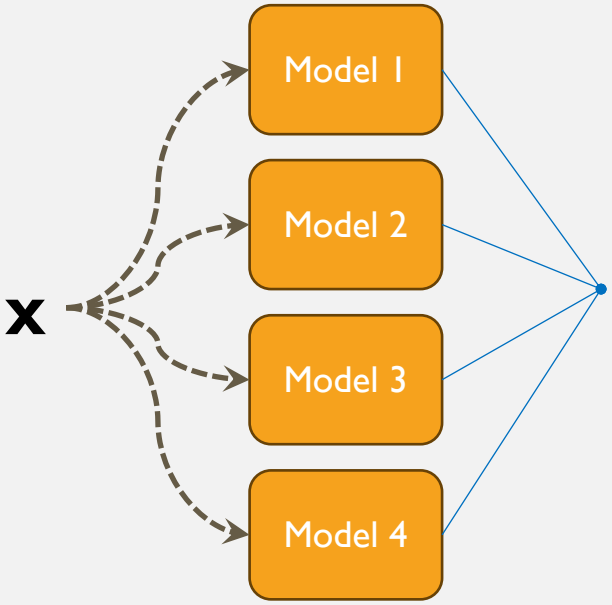
expected risk **(single model)** = noise + bias + variance

expected risk **(ensemble)** = noise + **average bias** + **average variance** − diversity

THE SHORT STORY

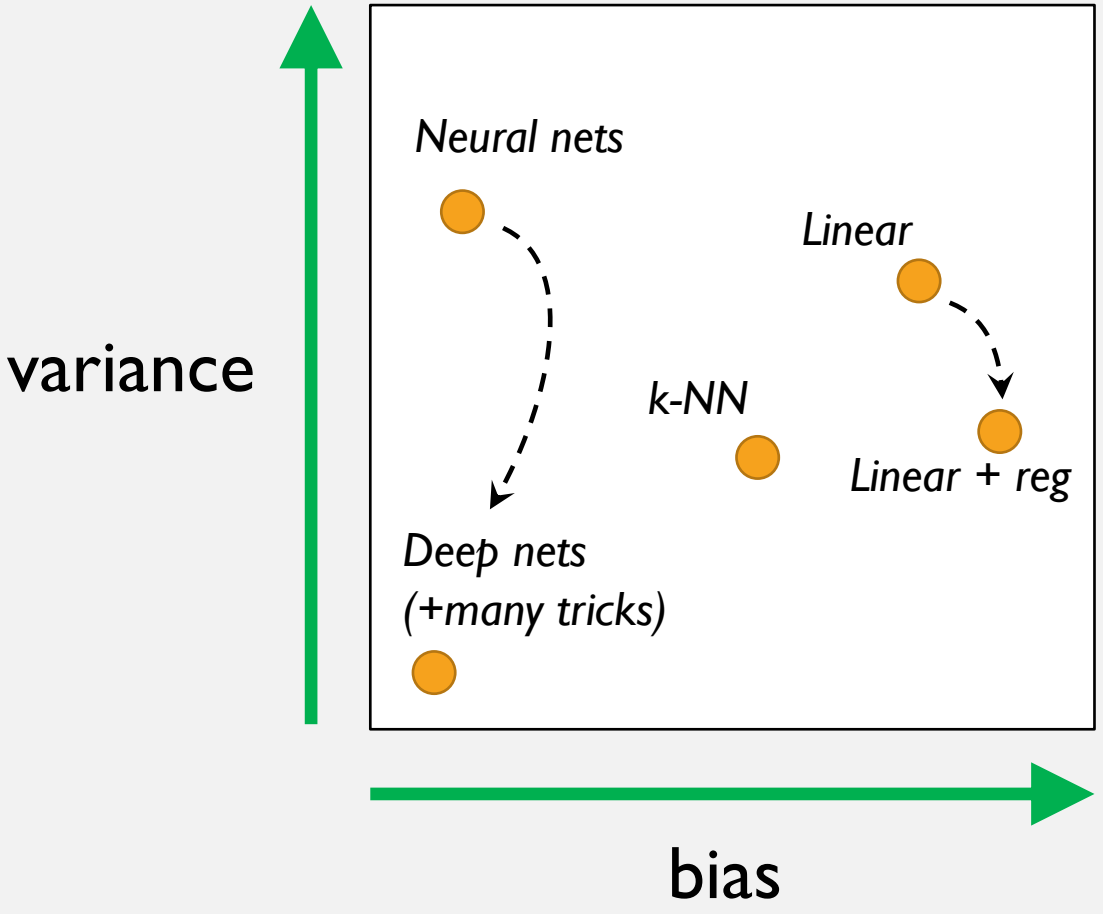
expected risk **(ensemble)** = noise + average **bias** + average **variance** − diversity

constant | strength of models | sensitivity of models | differences *between* models

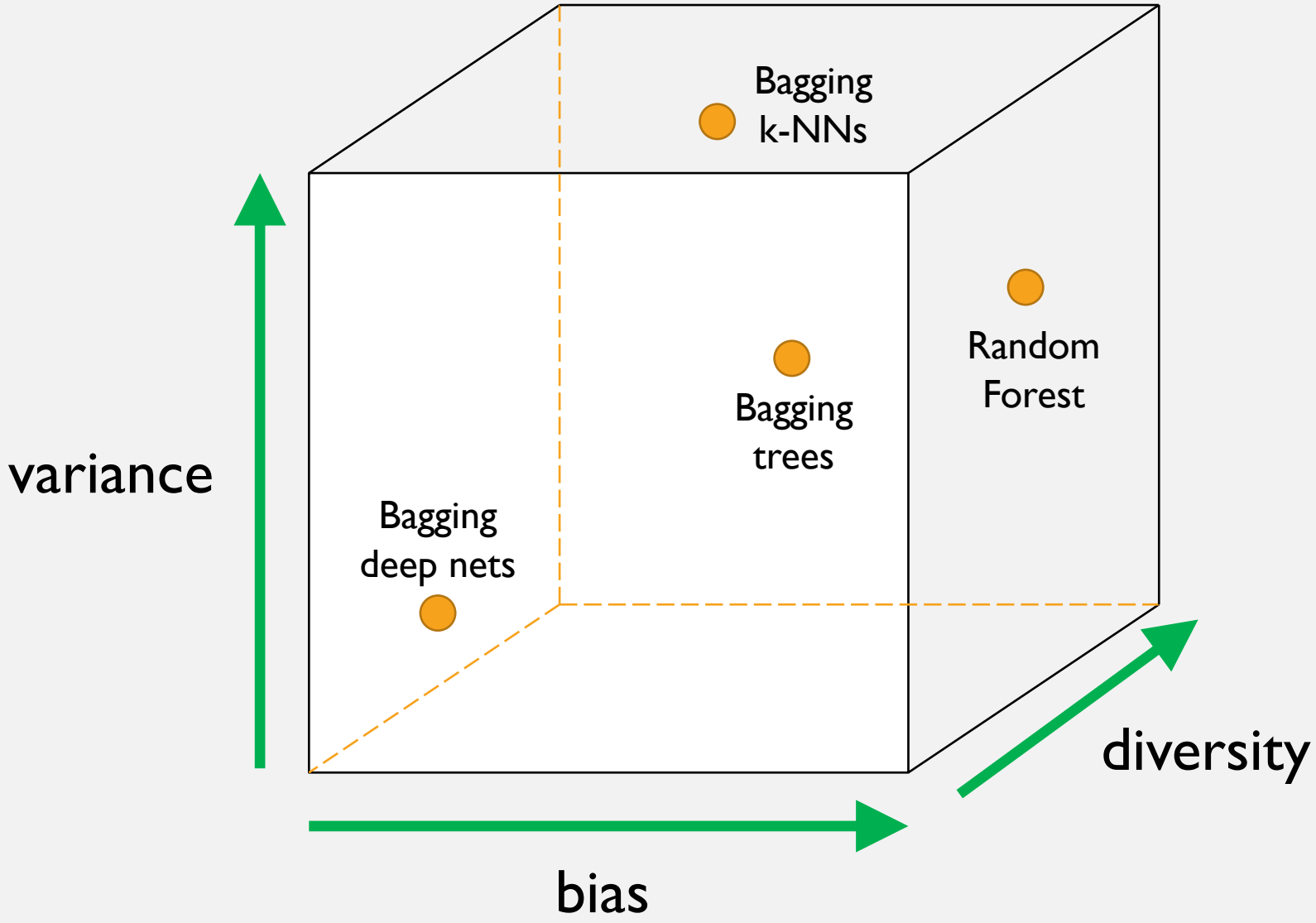


$$\text{diversity} = \mathbb{E}_D\left[\frac{1}{m}\sum_{i=1}^{m}(f_i(\mathbf{x}) - \bar{f}(\mathbf{x}))^2\right]$$

THE SHORT STORY: SINGLE MODELS

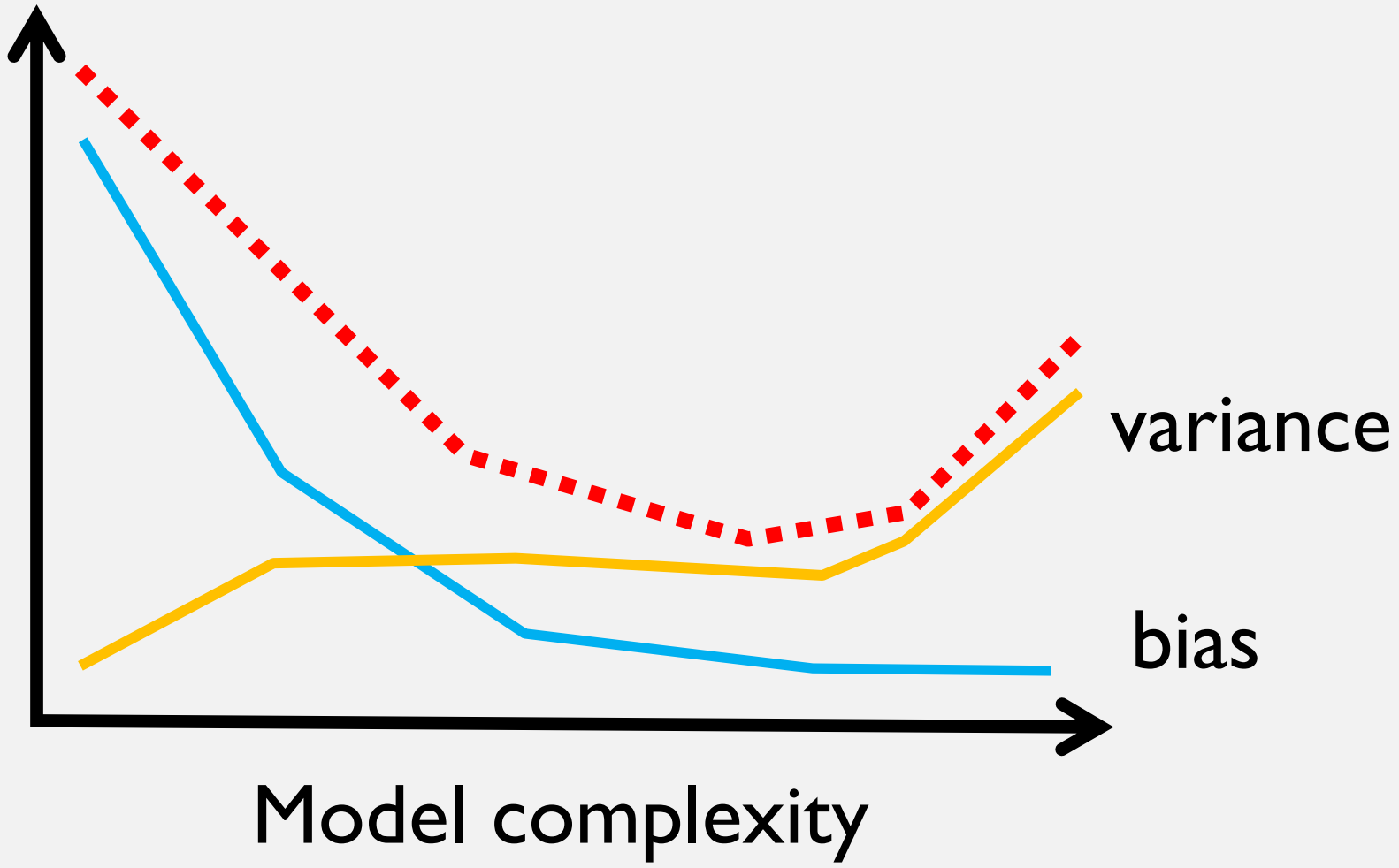


THE SHORT STORY: ENSEMBLE MODELS



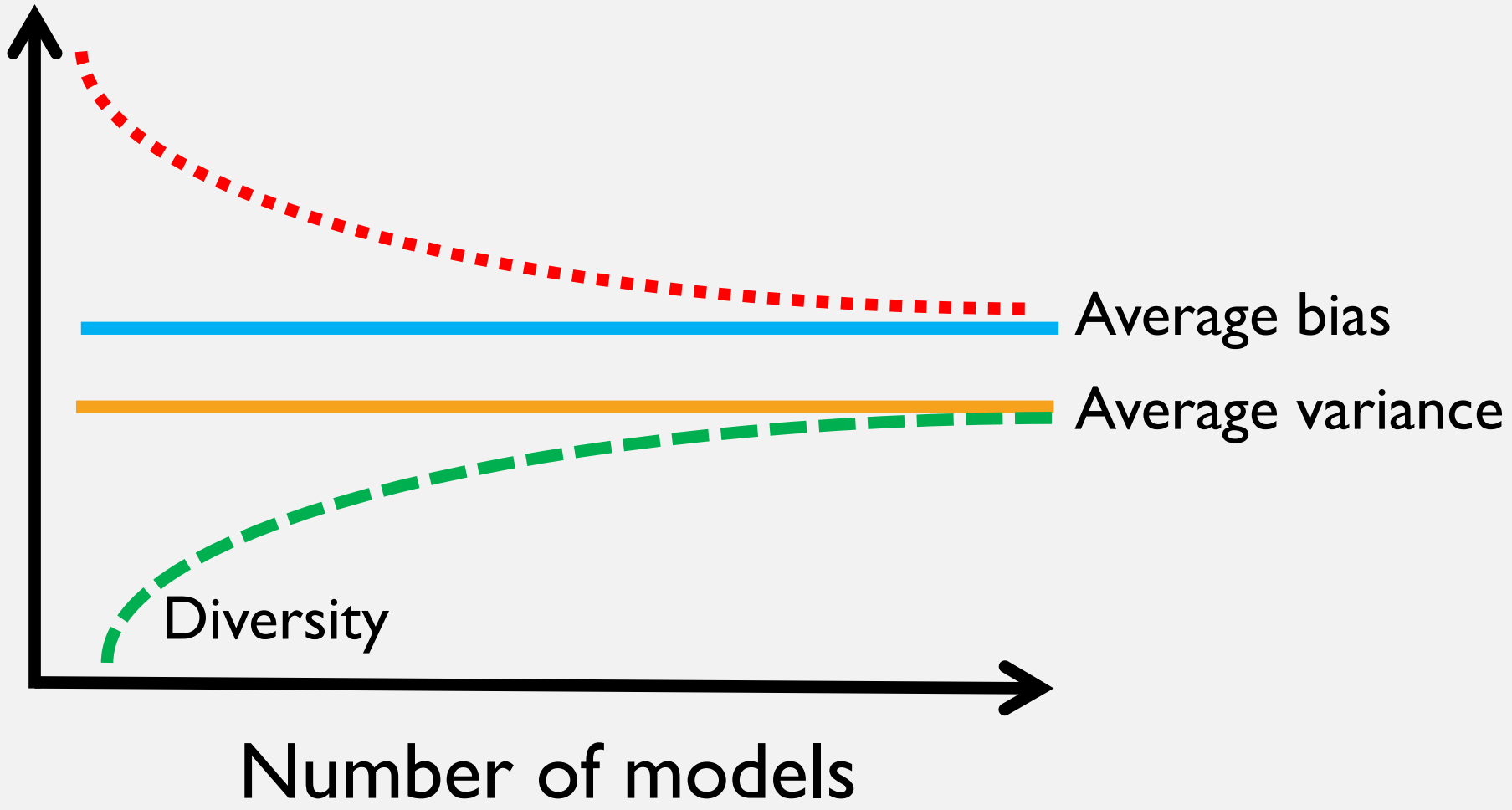
WHY DOES IT MATTER? OVERFITTING, THAT'S WHY.

A single model….



WHY DOES IT MATTER? OVERFITTING, THAT'S WHY.

An ensemble of decision trees, all of depth 8….



SO DIVERSITY APPLIES… WHEREVER BIAS/VARIANCE APPLIES

Expected ensemble loss = Average Bias + Average Variance - Diversity

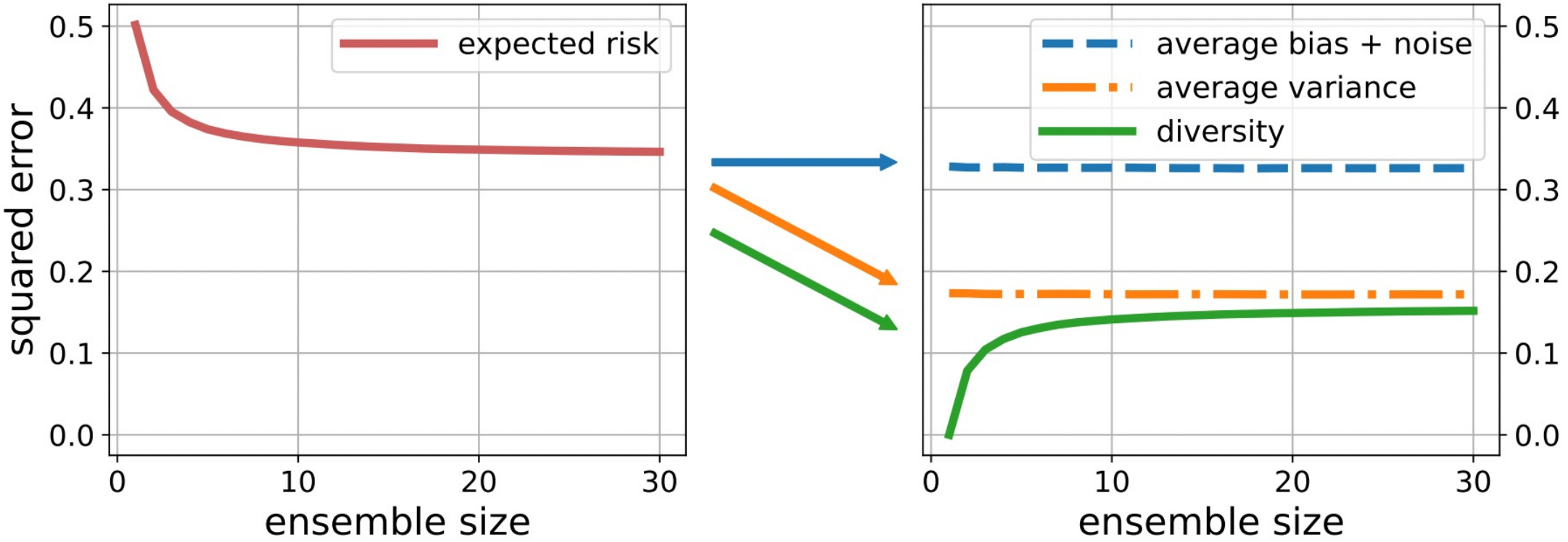


Figure 7: Decomposing the expected ensemble loss (Bagging depth 8 regression trees).

SO DIVERSITY APPLIES… WHEREVER BIAS/VARIANCE APPLIES

Expected ensemble loss = Average Bias + Average Variance - Diversity

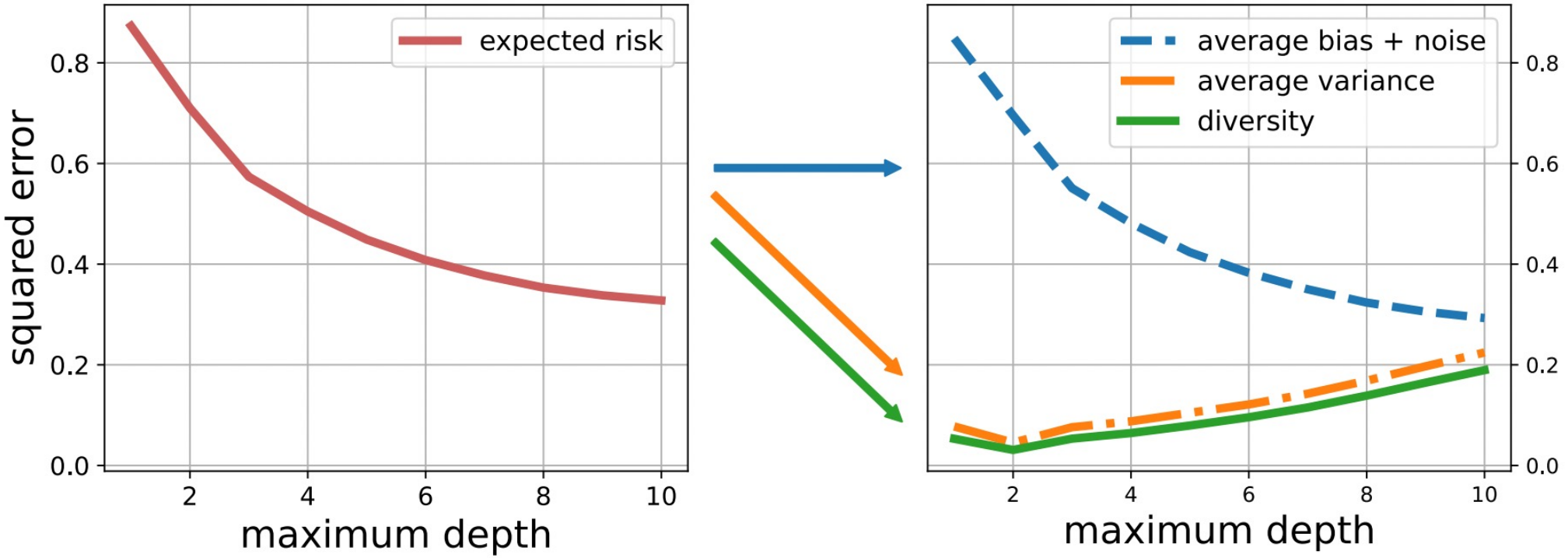


Figure 8: Bagging $M = 10$ trees, varying maximum depth.

THE MATHEMATICS

expected risk **(ensemble)** = noise + **average bias** + **average variance** − diversity

This is formalized as Theorem 5 in your assigned reading….

Theorem 5 (Generalized Bias-Variance-Diversity decomposition) *Consider a set of models* $\{\mathbf{q}_i\}_{i=1}^m$*, evaluated by a loss* ℓ*. Assuming a bias-variance decomposition holds in the form of Definition 2, the following decomposition also holds.*

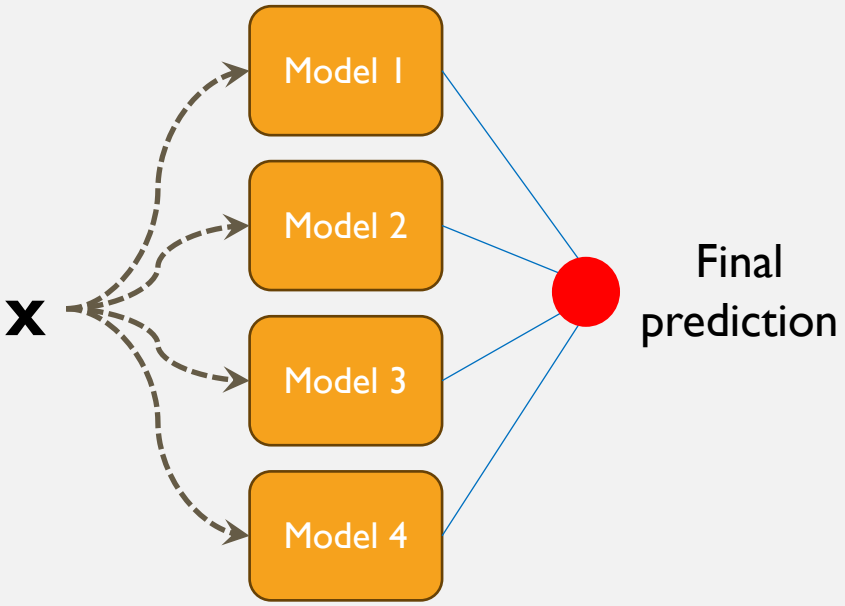
$$\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{XY}}[\ell(\mathbf{Y},\overline{\mathbf{q}})]\Big] =$$

$$\mathbb{E}_{\mathbf{X}}\Bigg[\underbrace{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\ell(\mathbf{Y},\mathbf{Y}^*)]}_{\text{noise}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{Y}^*,\mathring{\mathbf{q}}_i)}_{\text{average bias}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[\ell(\mathring{\mathbf{q}}_i,\mathbf{q}_i)\right]}_{\text{average variance}} - \underbrace{\mathbb{E}_D\left[\frac{1}{m}\sum_{i=1}^m \ell(\overline{\mathbf{q}},\mathbf{q}_i)\right]}_{\text{diversity}}\Bigg], \quad (9)$$

where $\mathring{\mathbf{q}} \stackrel{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \mathbb{E}_D\Big[\ell(\mathbf{z},\mathbf{q})\Big]$ *and the combiner is* $\overline{\mathbf{q}} \stackrel{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \frac{1}{m}\sum_{i=1}^m \ell(\mathbf{z},\mathbf{q}_i)$.

THE MATHEMATICS

expected risk **(ensemble)** = noise + **average bias** + **average variance** − diversity



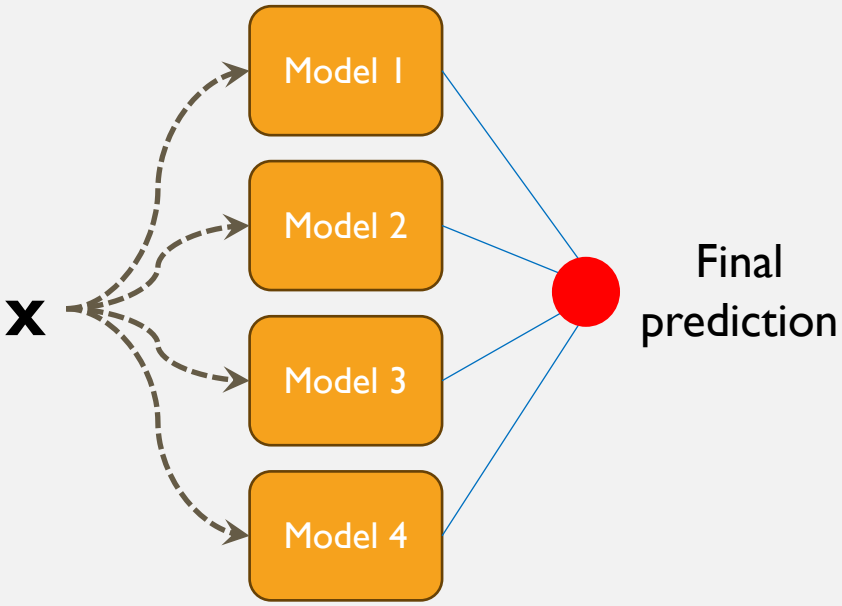
For this equality to hold….

… we must again use a particular combiner – the "centroid".

Loss function	Centroid Combiner	Name
Squared loss	$\frac{1}{m}\sum_{i=1}^{m} q_i$	Arithmetic mean
Poisson regression loss	$\prod_{i=1}^{m} q_i^{\frac{1}{m}}$	Geometric mean
KL-divergence	$Z^{-1}\prod_{i=1}^{m}\left(\mathbf{q}_i^{(c)}\right)^{\frac{1}{m}}$	Normalised geometric mean
Itakura-Saito loss	$1/\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{q_i}\right)$	Harmonic mean

CENTROID COMBINER FOR THE CROSS-ENTROPY

expected risk **(ensemble)** = noise + average **bias** + average **variance** − diversity



If our loss is cross-entropy:

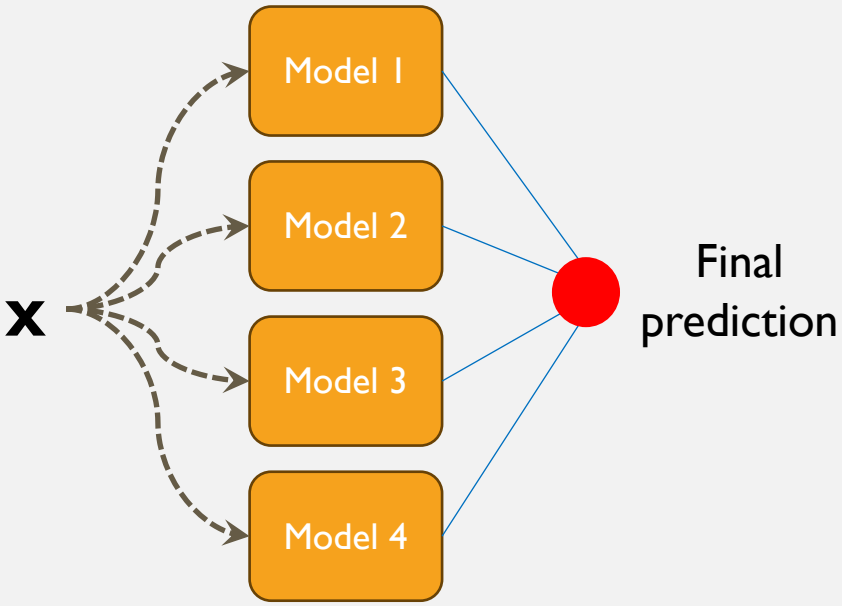
$$\ell(y, f(\boldsymbol{x})) \;:=\; -\Big[y \ln f(\boldsymbol{x}) + (1-y)\ln(1-f(\boldsymbol{x}))\Big]$$

The centroid is the normalized geometric mean…

$$\bar{f}(\mathbf{x}) = Z^{-1} \prod_i f_i(\mathbf{x})^{1/m}$$

NORMALIZED GEOMETRIC MEAN FOR CROSS-ENTROPY

$$\bar{f}(\mathbf{x}) = Z^{-1} \prod_i f_i(\mathbf{x})^{1/m}$$



	f_1	f_2	f_3	$\prod$	$\sqrt[3]{}$
Class 1	0.8	0.6	0.9	0.432	0.7559
Class 2	0.2	0.4	0.1	0.008	0.2

Z = 0.955952

Class 1	0.75595 / 0.955952	= 0.7907...	$\bar{f}(\mathbf{x})$
Class 2	0.2 / 0.955952	= 0.2092...	

CONCLUSION

We can make models "bigger" by making committees of them…

…but the same sort of rules apply.

In single models we have a 2-way tradeoff (**bias/variance**).

In ensembles of models, it's a a 3-way tradeoff (**bias/variance/diversity**).

…. But… it only holds if we use the centroid combiner rule

expected risk **(ensemble)** = noise + average **bias** + average **variance** − diversity