Post-Regularization: a Systematic Taxonomy

a seminar presentation by Yara Mohammadi Bahram

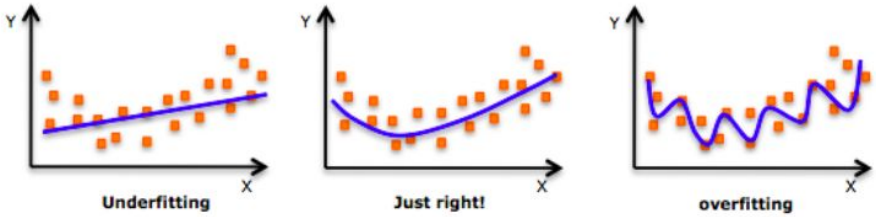
Supervisor: Dr. Amin Sadeghi

Presentation Overview

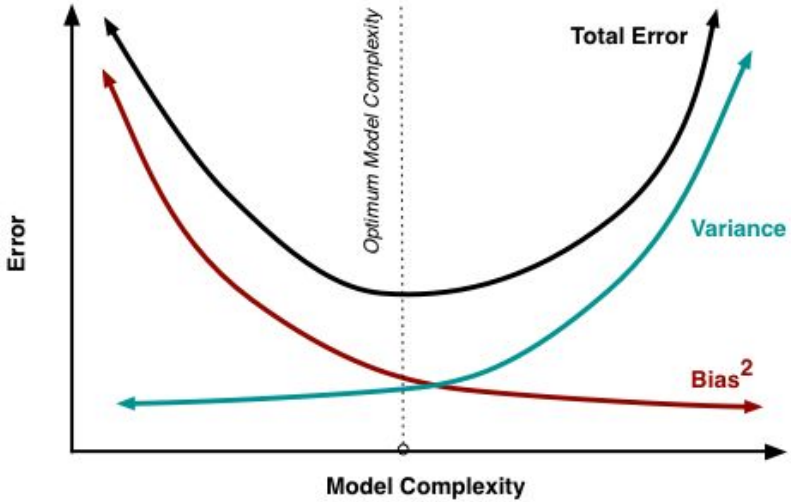
- Motivation
  - What is regularization?
  - Problems of conventional regularization
  - What is post-regularization
- Taxonomy
  - Top-level overview
    - Early regularization
    - Late regularization
    - Post-regularization
- Final remarks

Motivation

What is Regularization?



(Kukaˇcka et al., 2017)
“Any method of variance reduction form that aims at making the model generalize better, i.e., produce better results on the test set”.



Problems of conventional regularization

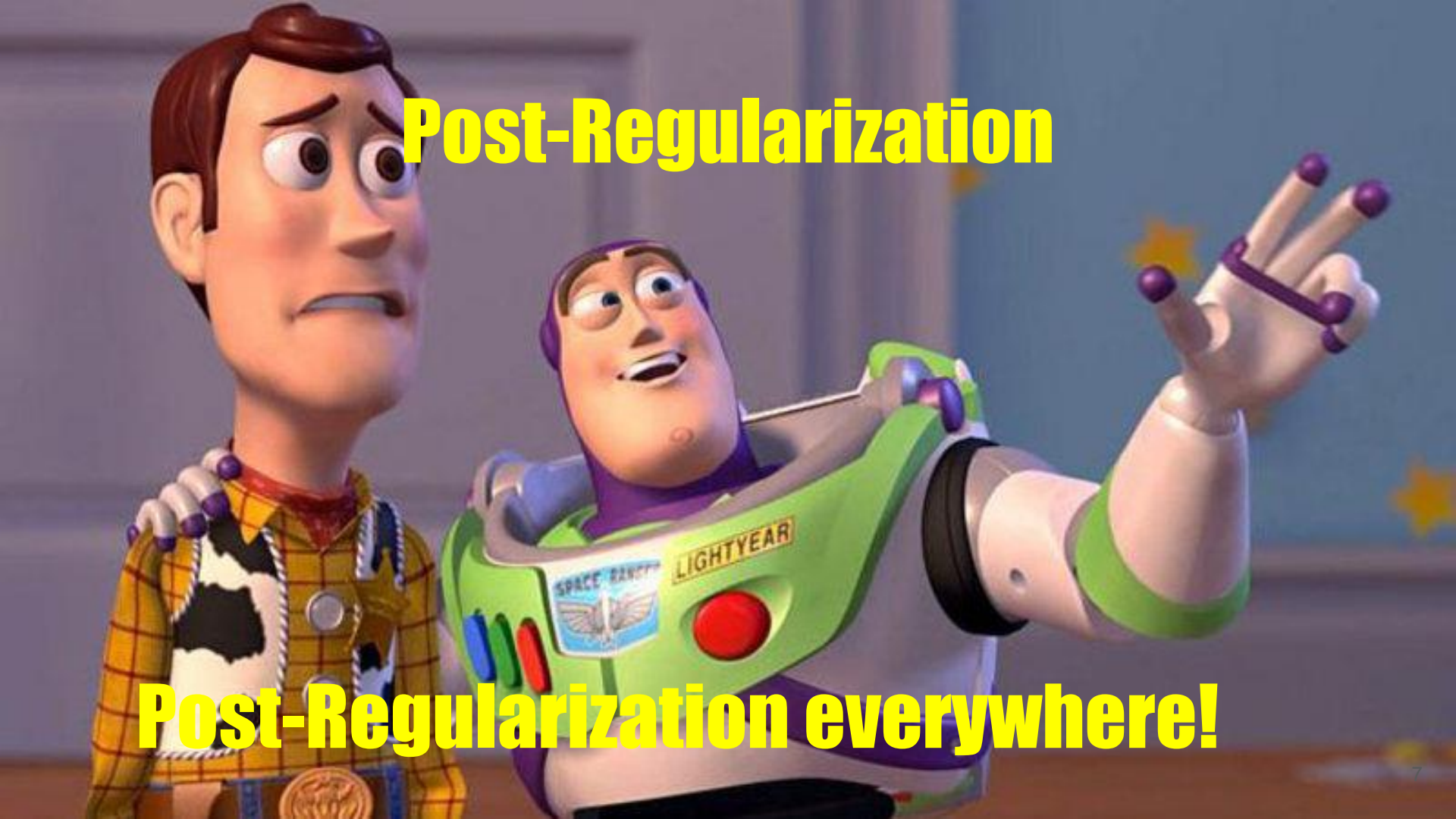
- Regularization methods are traditionally integrated into the training process
  - modifying the strength of regularization on a trained model usually requires retraining the model from scratch.

- Deep learning models are becoming more and more complex
  - The use of pre-trained models is becoming more common

…Disentangling the regularization process from model training is an increasingly viable concept that has been almost neglected so far.

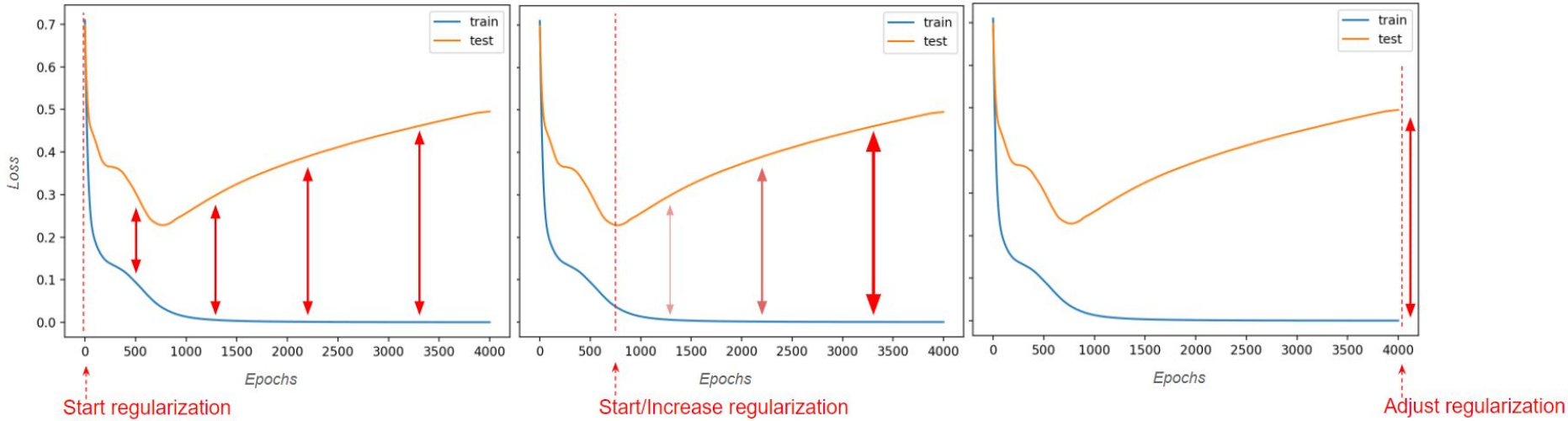
What is Post-Regularization?

- Regularization methods, that are disentangled from the model training process
- regularization methods, that are specifically crafted to take pre-trained unregularized models out of their overfitting solutions

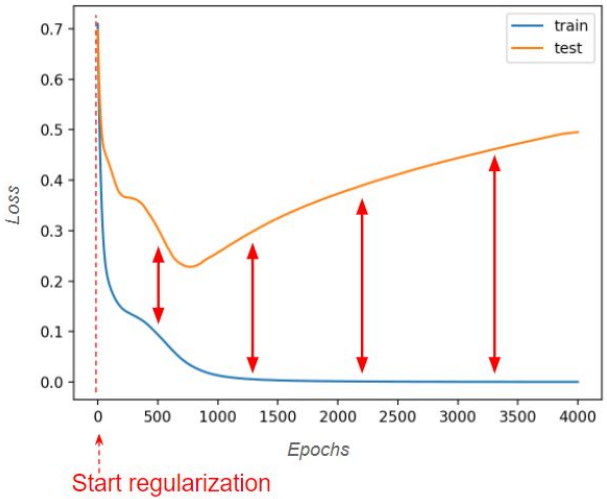

Post-Regularization
LIGHTYEAR
Post-Regularization everywhere!

Taxonomy

Top-level overview



Early regularization



A mini-taxonomy for conventional regularizations

$$Minimize_w \frac{1}{|D|} \sum_{(x_i,y_i)\in D} L(f_w(x_i), y_i) + R(...)$$

Regularization via:

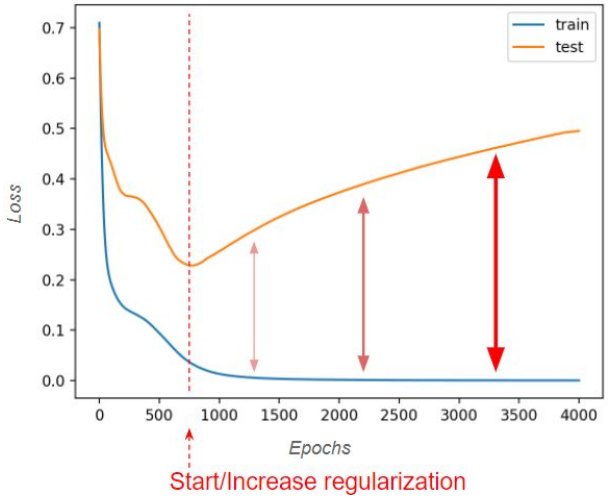
- Data
- Model
- Error function
- Regularization term
- Optimization

Early regularization

Type	Method family	Method	Papers	Description
Fixed	Via data	Adding noise, Batch normalization, Dropout, Augmentation, Data synthesis, Adversarial training, Label smoothing, Model compression, ...	(Kukačka et al., 2017), (Moradi et al., 2020)	Conventional regularizations that usually are applied as a part of the training and cannot be modified after the training.
	Via Model	Choice of architecture, layers, activation functions, model selection, multi-task learning, Ensemble, ...		
	Via loss function	Weight decay, Jacobian, Similarity measures, ...		
	Via Optimization	Weight smoothing, Flat minima search, Hessian penalty, Choice of optimization, Momentum, Learning rate, ...		
Adjustable	*	Dynamic Regularization	(Wang et al., 2020)	Modeling the regularization strength as a function of the training loss in a CNN.
Reducing search cost	*	Neural Architecture search, Hyperparameter optimization	(Ren et al., 2021), (Yu & Zhu, 2020)	Improving the time needed to find the optimal hyperparameters for the model and task.

Table 1: Methods of early regularization (Section 3).

Late regularization

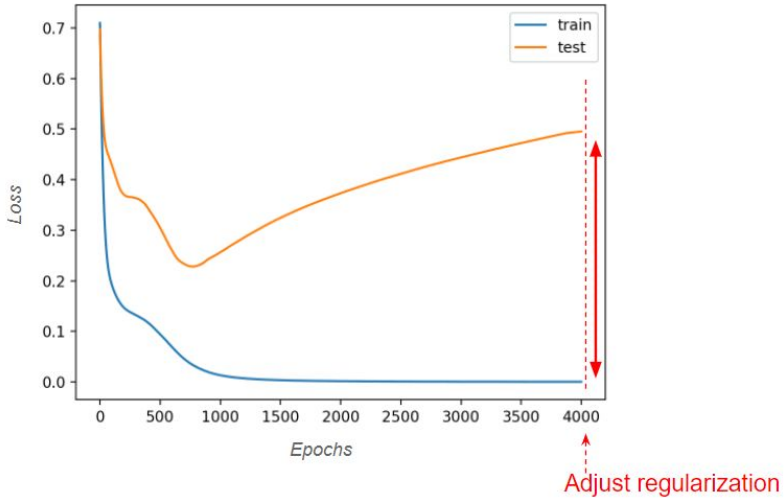


Late regularization

Type	Method	Papers	Description
Increase (Curriculum regularization)	Curriculum Magnitude Pruning	(Bartoldson et al., 2019)	Introducing and increasing pruning strength in later epochs.
	Curriculum dropout	(Morerio et al., 2017)	Increasing dropout strength in later epochs
	Efficientnet V2	(Tan & Le, 2021)	Increasing the size of input images and augmentation strength in later epochs.
Start (fine-tuning)	Self-distillation	(Zhang et al., 2019), (Mobahi et al., 2020)	Smoothening the output and sparsifying weights of a pre-trained model by self-distillation post-training
	Adding a regularization term	(Jakubovitz & Giryes, 2018)	Improving generalization and robustness by post-training with an additional Jacobian regularization term
	Feature augmentations	(Khan & Fraz, 2020), (Kapoor et al., 2020)	Fine-tuning neural networks with additional self-supervision using feature augmentations
	Retraining the last layer	(Moreau & Audiffren, 2016)	Showing that splitting the training to representation learning and fine-tuning the last layer improves generalization
	Unsupervised fine-tuning	(Cerisara et al., 2021)	Optimizing the approximation of true classifier risk after supervised training. (Converges to a wider region)
	Post-training quantization and pruning	(Lazarevich et al., 2021) (Gholami et al., 2021)	Compressing pre-trained models.
	Dropout for transfer-learning	(Wu et al., 2021)	Forcing the output distributions of different sub-models generated by dropout to be consistent with each other
	Transfer learning in Neural Machine Translation	(Barone et al., 2017)	Exploring variants of L2 and Dropout regularization for domain-adaptation of pre-trained NMT models
	Regularizing transfer-learning	(Li & Zhang, 2021)	Improving regularization and robustness by Fine-tuning neural networks by L2 norm and label weight correction regularization

Table 2: Methods of late regularization (Section 4). Fine-tuning regularizations can fit inside the post-regularization table as well.

Post-regularization



A demonstration

- Simple polynomial regression:

$$f(x) = c_0 + c_1 x + c_2 x^2 \ldots c_n x^n$$

$(1, x, x^2, x^3)$ Multicollinearity, multicollinearity everywhere…

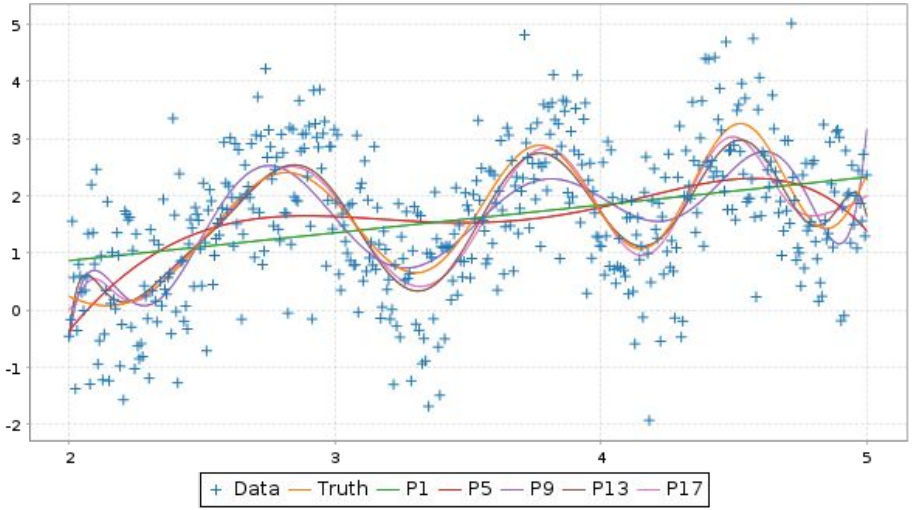
- Orthogonal polynomial regression:

$$P_0(x_i) = 1$$

$$P_1(x_i) = \lambda_1 \left[\frac{x_i - \bar{x}}{d} \right]$$

$$P_2(x_i) = \lambda_2 \left[\left(\frac{(x_i - \bar{x})}{d} \right)^2 - \left(\frac{n^2 - 1}{12} \right) \right]$$

$$P_3(x_i) = \lambda_3 \left[\left(\frac{(x_i - \bar{x})}{d} \right)^3 - \left(\frac{x_i - \bar{x}}{d} \right) \left(\frac{3n^2 - 7}{20} \right) \right]$$



Post-regularization

Type	Method family	Method	Papers	Description
Modified Training	Via disentangling	PCA, Fourier transform, JPEG	-	Impose orthogonality in feature/function vectors
		VAE, Orthogonal DNN	(Kingma & Welling, 2019), (Choi et al., 2020)	Disentangle features for easier feature selection
		Continual Learning	(Mai et al., 2022)	Disentangling what the model learns across different augmentation and regularization strengths.
		Quantization and Pruning aware training	(Gholami et al., 2021)	Make model more flexible for feature/model selection
		Slimmable neural networks	(Yu & Huang, 2019b), (Yu & Huang, 2019a)	Training networks in a way to be executable at arbitrary width with instant and adaptive accuracy-efficiency trade-off at runtime
	Via ensembling	Iterate Averaging	(Izmailov et al., 2018), (Wu et al., 2020), (Granziol et al., 2020)	Saving the SGD optimization path and adjusting the regularization strength after training via geometric averaging of the path ensemble.
		Bagging, Gradient boosting, ResNet	(Siu, 2019)	Training a series of weak classifiers and deciding how many of them to keep at runtime. ResNet behaves like boosting algorithms
Black-box	*	Certified robustness	(Cohen et al., 2019)	Certified adversarial robustness via Randomized smoothing
		Test-time augmentations	(Kim et al., 2020), (Cohen & Giryes, 2021) (Shanmugam et al., 2021)	Averaging the prediction over different transformations of the input data at test-time

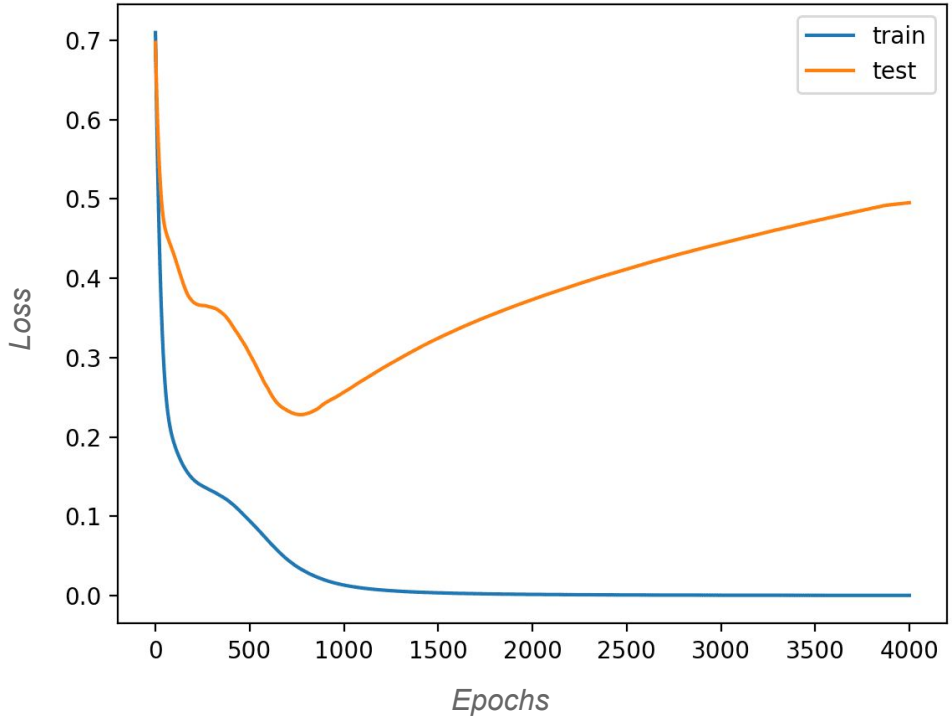
Table 3: Methods of post-regularization (Section 5).

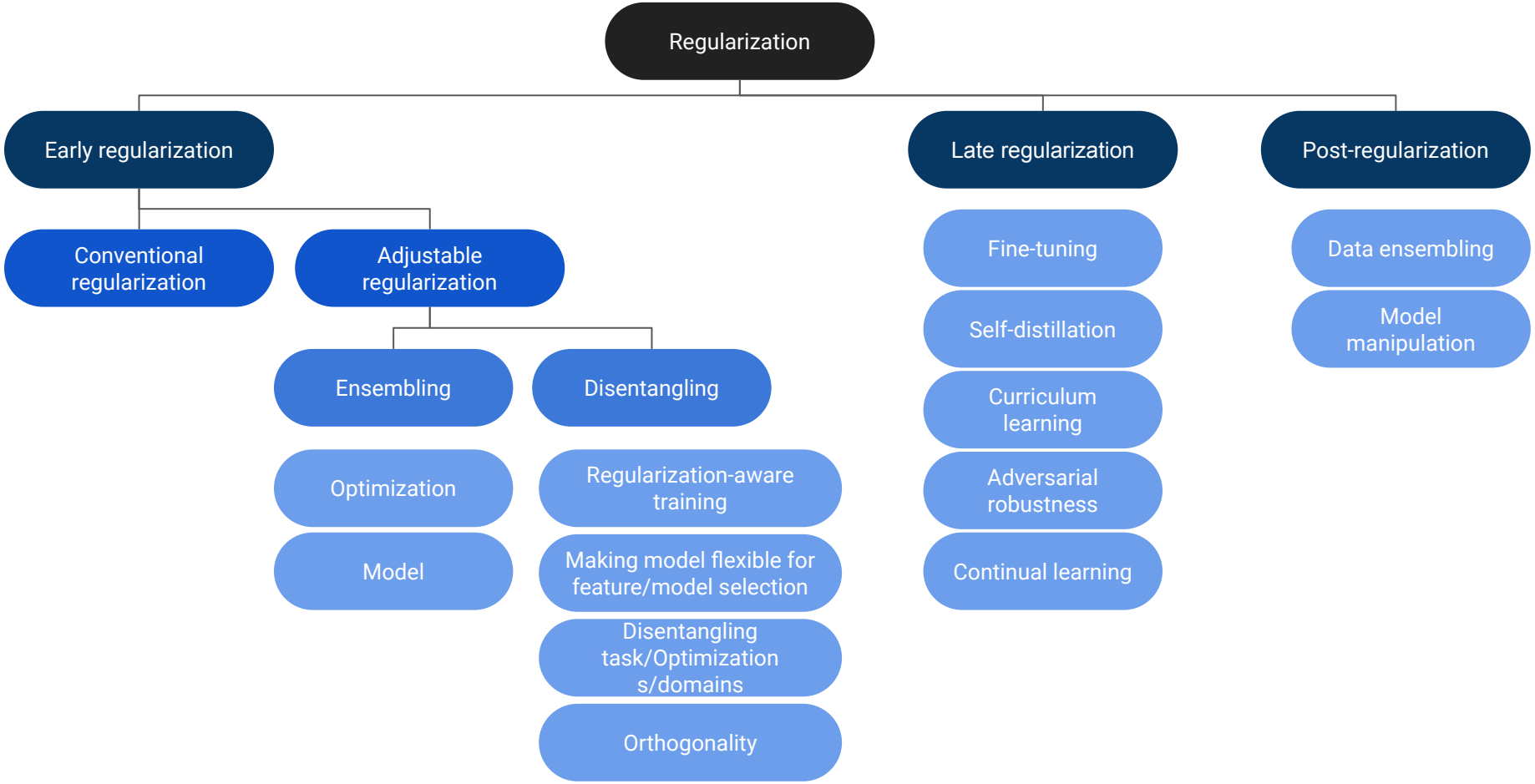
Final remarks

- Post-Regularization is an increasingly viable concept in machine learning and It requires more attention.
- We hope to enable the discovery of new and improved post-regularization methods by our taxonomy.


Thank you

No regularization (Overfit)




Regularization
Early regularization
Late regularization
Post-regularization
Conventional regularization
Adjustable regularization
Fine-tuning
Self-distillation
Curriculum learning
Adversarial robustness
Continual learning
Data ensembling
Model manipulation
Ensembling
Disentangling
Optimization
Model
Regularization-aware training
Making model flexible for feature/model selection
Disentangling task/Optimizations/domains
Orthogonality


Validation
On the same domain
On a different domain/task
On all domains
Fine-tuning
Improving generalization
Avoid overfitting
Battle overfitting
Transfer Learning
Domain Adaptation
Avoid overfitting on new domain
Continual learning
Few-shot learning
Domain generalization
Meta Learning
Lifelong learning
Flexible Domain
Editable neural networks
Reinforcement learning