Prediction Spatio-temporal Instability: Unraveling the Mechanism of Epoch-wise Double Descent

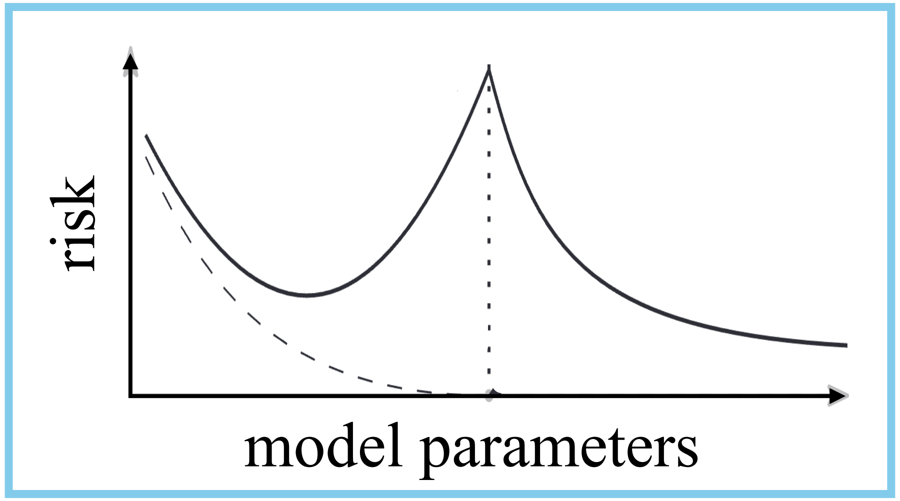
Keito Kobayashi， Reiya Hiramoto， Ryouichi Sekiguchi， Eisaku Maeda

Tokyo Denki University

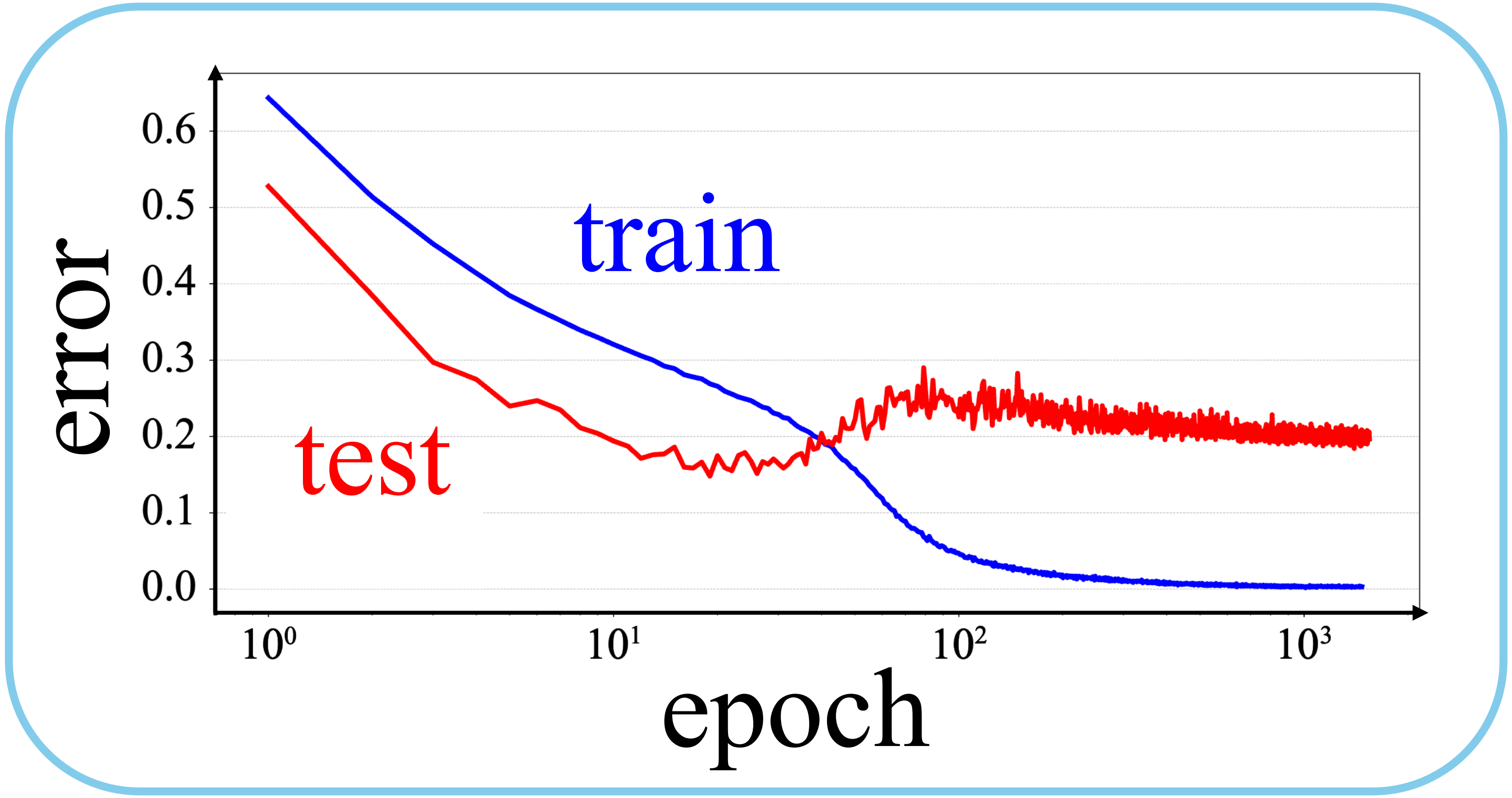
TOKYO DENKI UNIVERSITY

Double Descent

Model-wise Double Descent (MDD) [1]



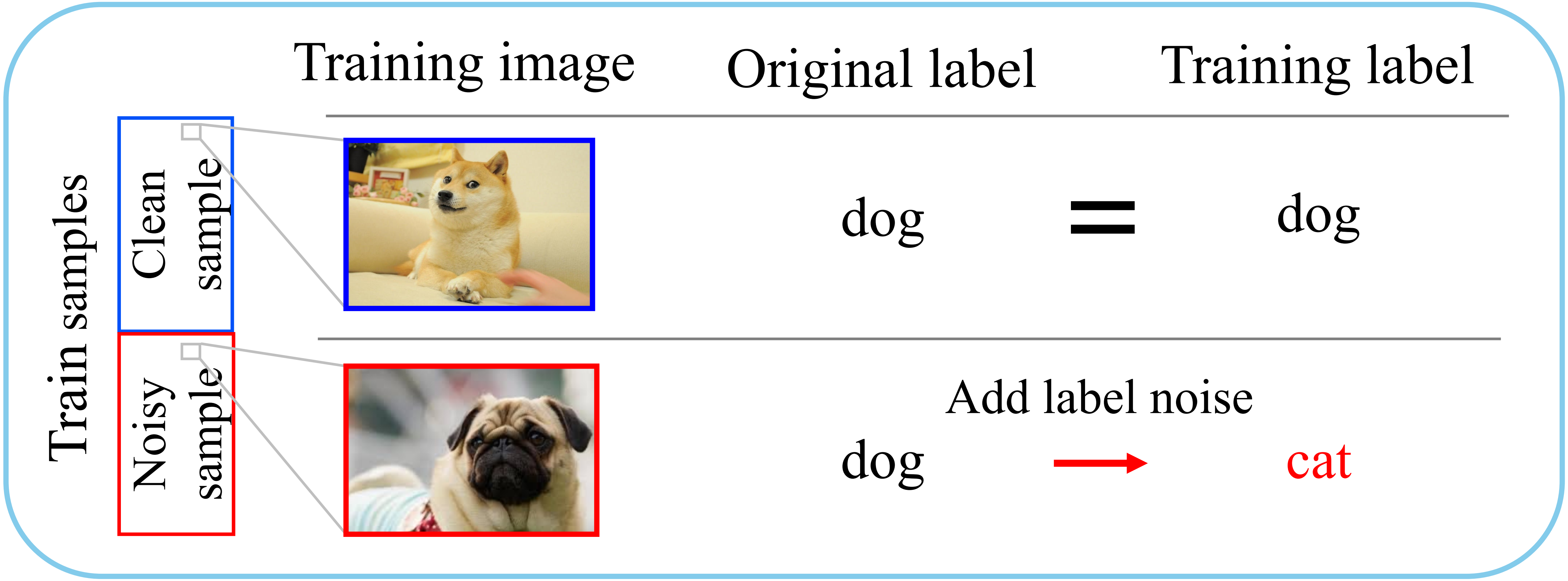
Epoch-wise Double Descent (EDD) [2]



CIFAR10, ResNet-18, with 20% label noise

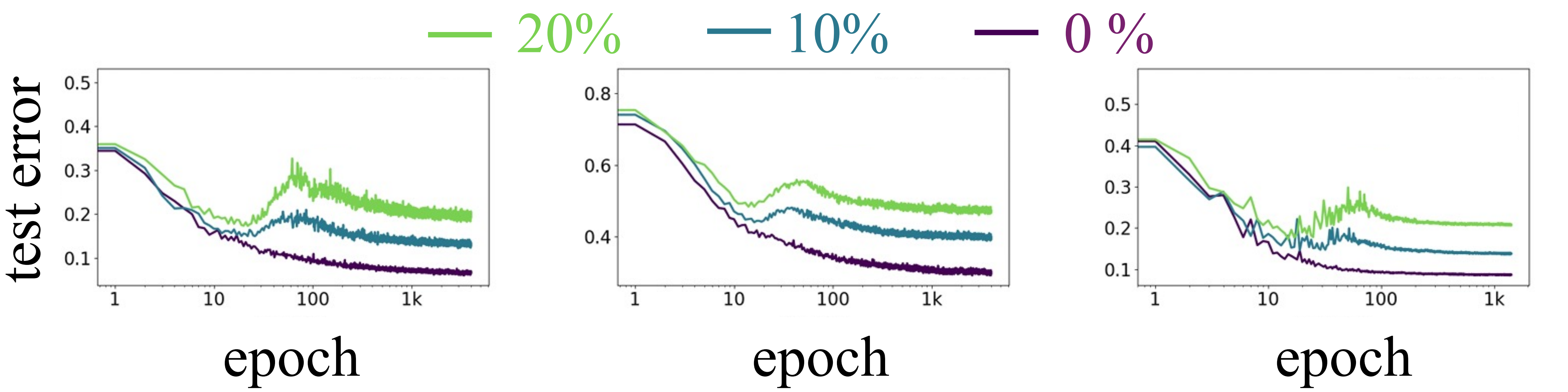
[1] Belkin+（PNAS2019）, [2] Nakkiran+（ICLR2020）

Label Noise



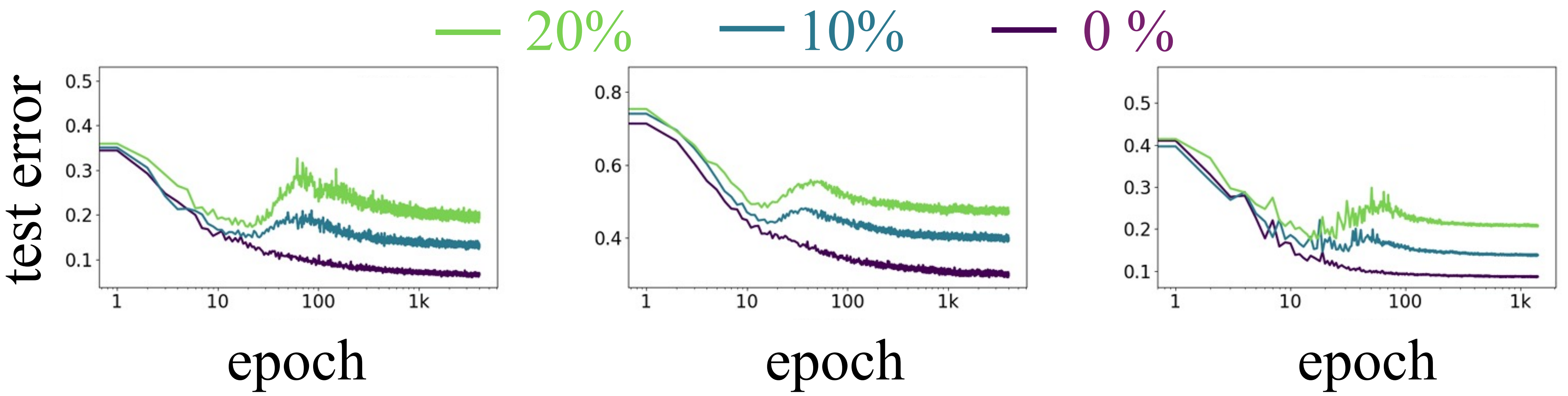
Training image
Original label
Training label
Train samples
Clean sample
dog
=
dog
Noisy sample
Add label noise
dog
cat

Epoch-wise DD was not observed at **0% label noise**



ResNet18 on CIFAR 10 ResNet18 on CIFAR 100 CNN 5layer on CIFAR 10

Epoch-wise DD was not observed at **0% label noise**



ResNet18 on CIFAR 10 | ResNet18 on CIFAR 100 | CNN 5layer on CIFAR 10

Investigate how noisy samples affect epoch-wise DD

Our Approach

- **Separate error-rate** analysis for **clean and noisy samples**

- Analyze the effect of noisy samples through
  spatio-temporal instability of predictions in the input space

Experiment

model

CNN 5layer

learning rate　0.01

optimizer　SGD

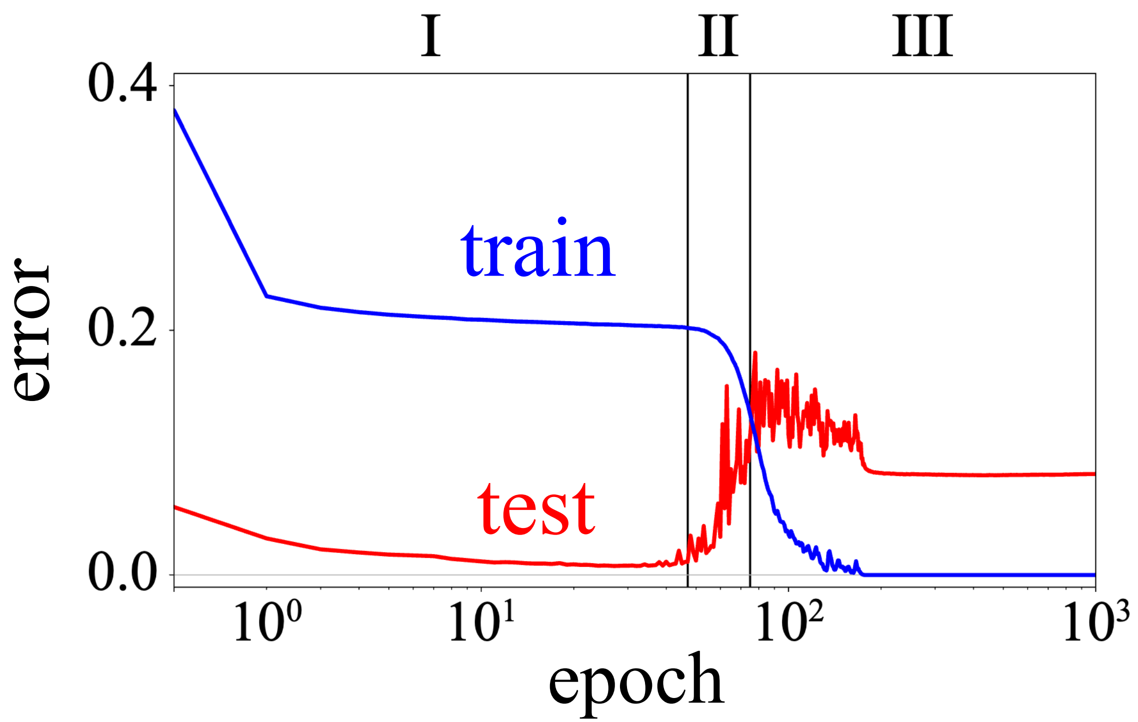
dataset

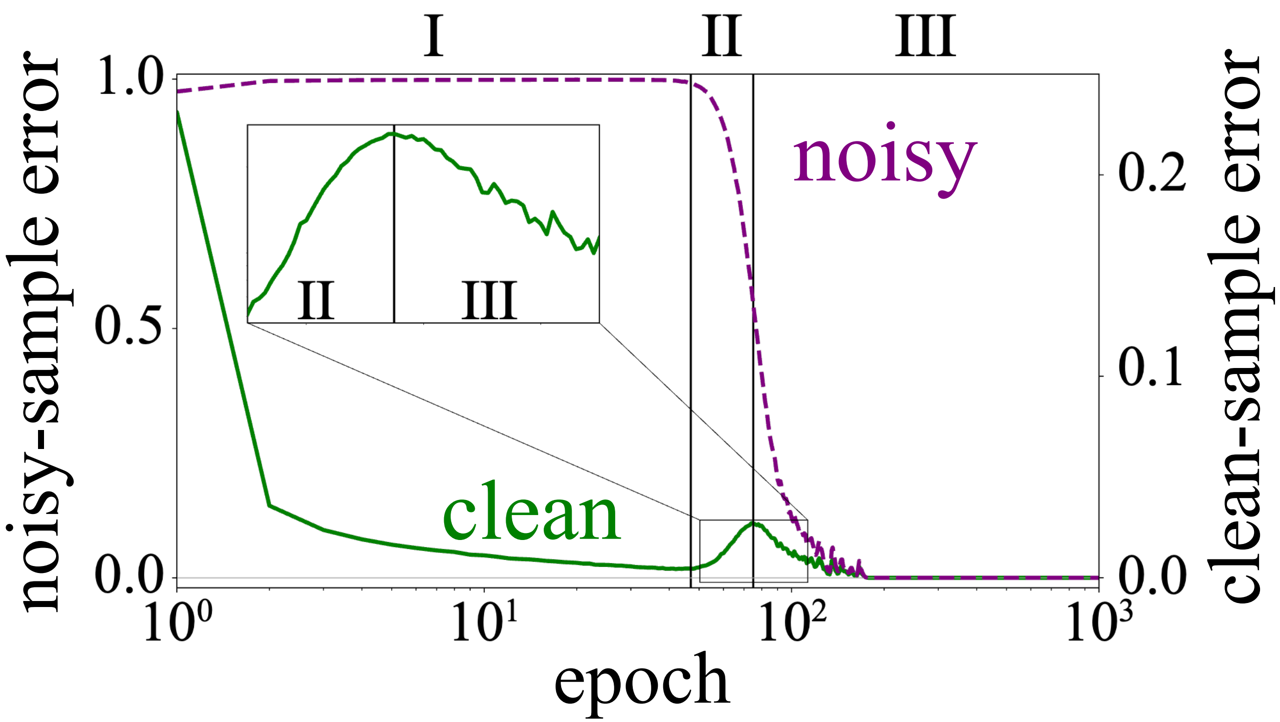
EMNIST Digits [4]

with 20% label noise

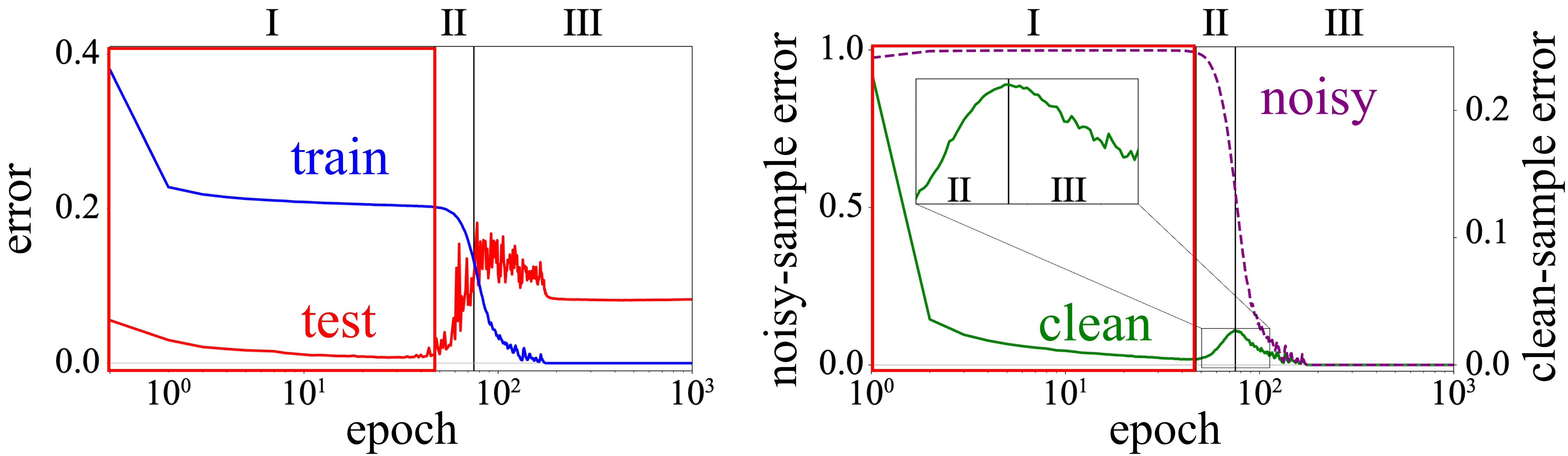
[4] Cohen+（IJCNN2017）

Learning Curve

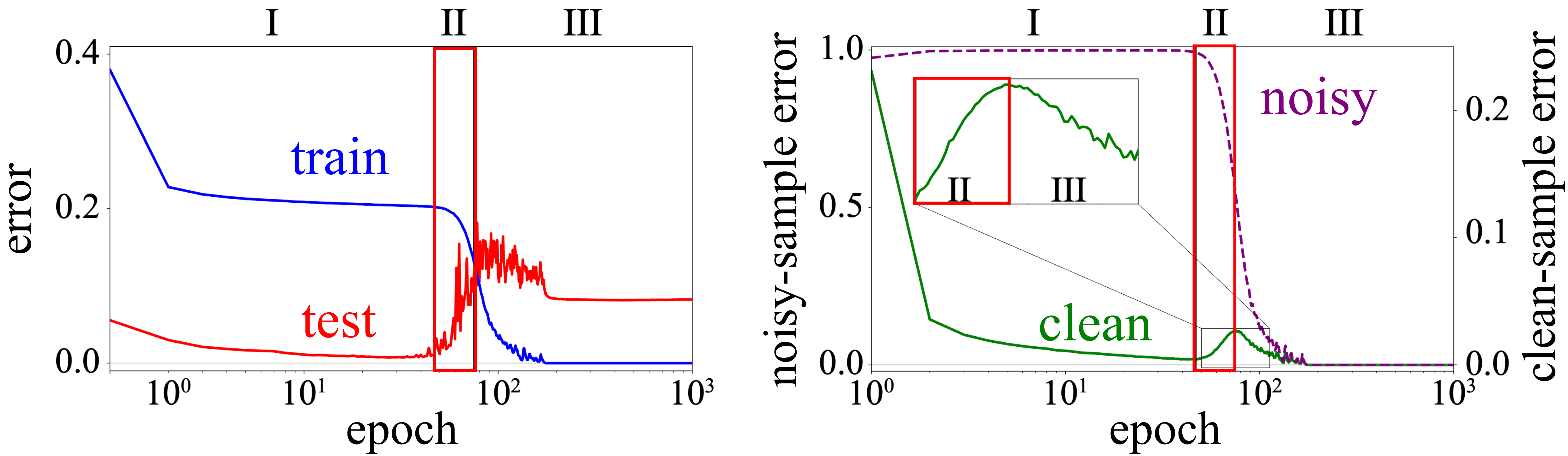




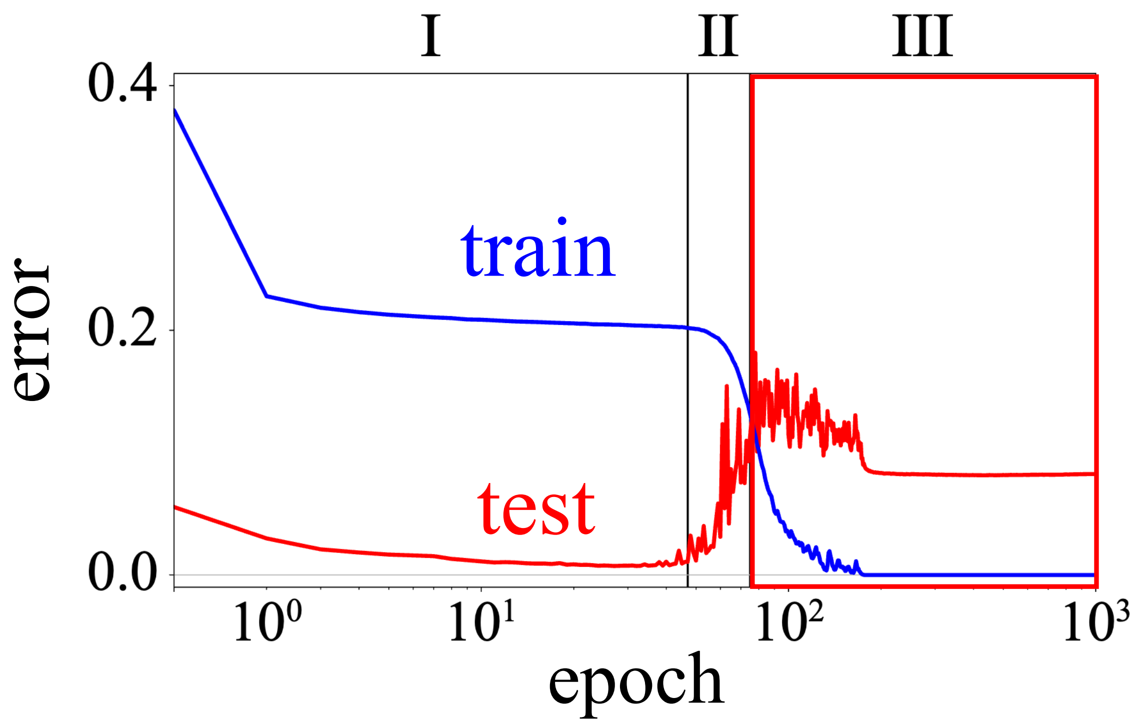
Learning Curve

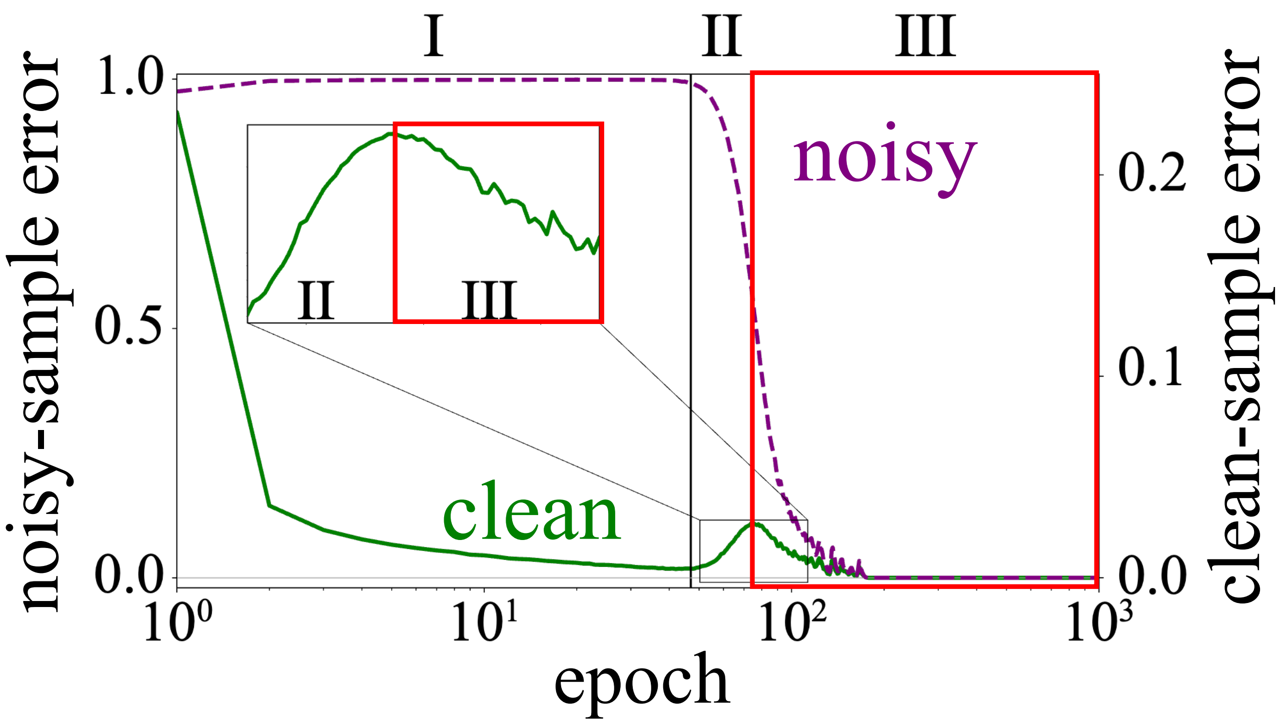


Learning Curve

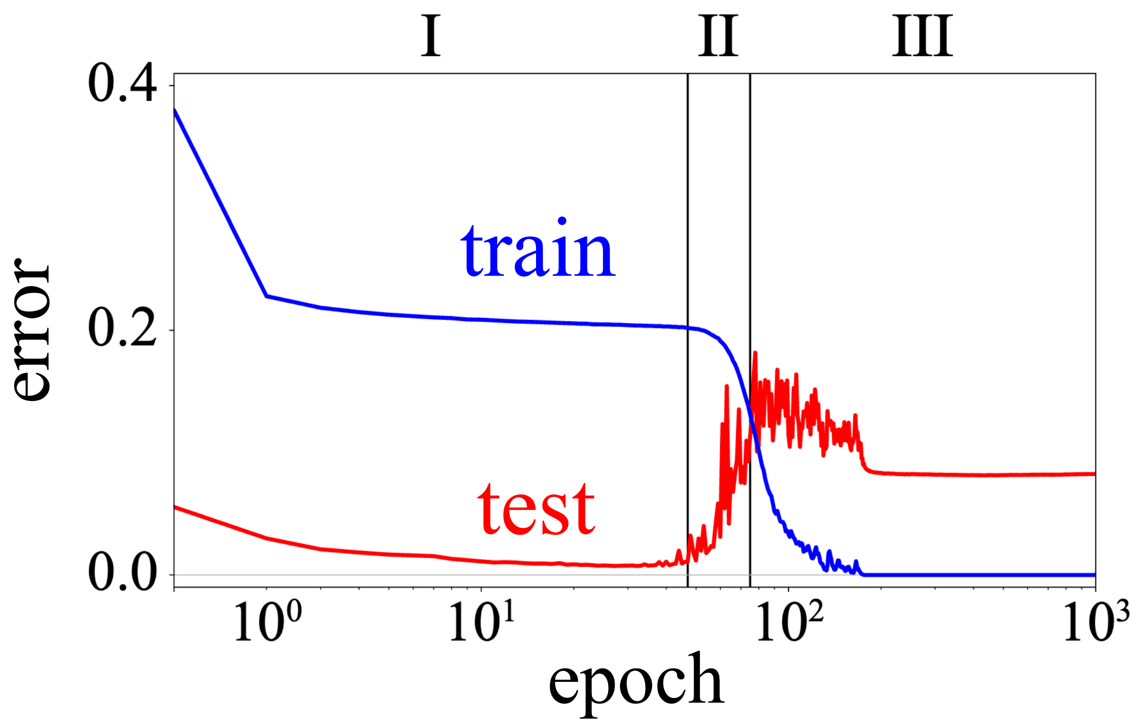


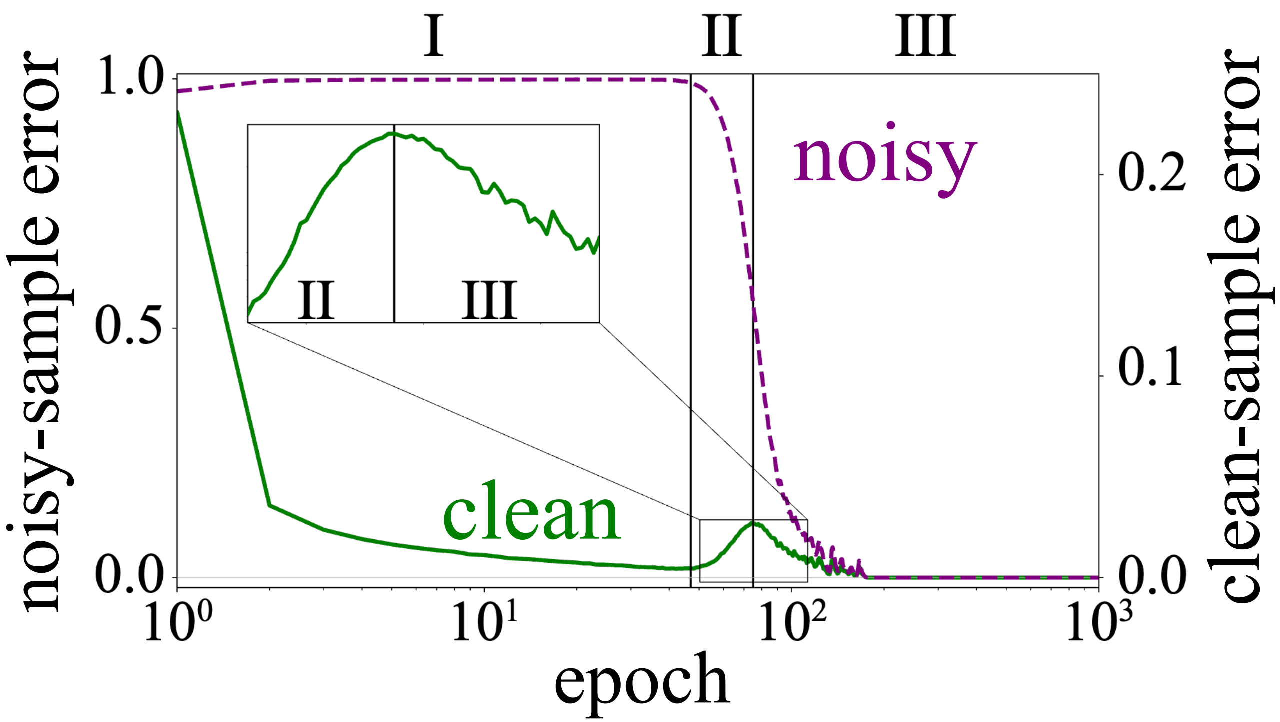
Learning Curve





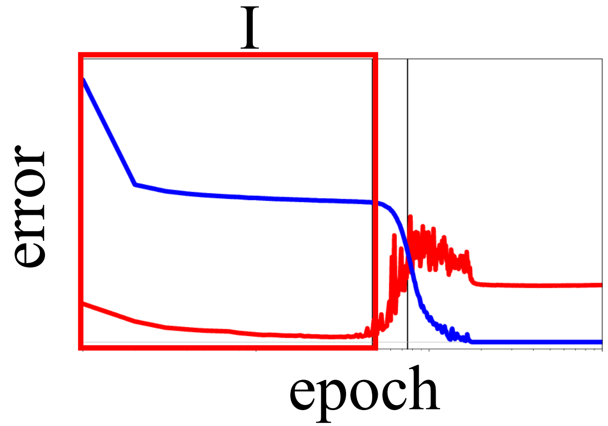
Learning Curve

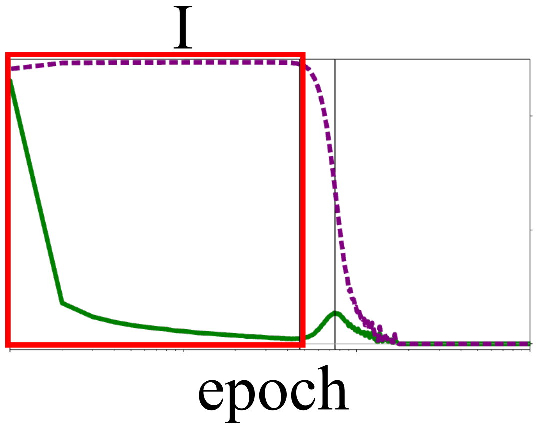


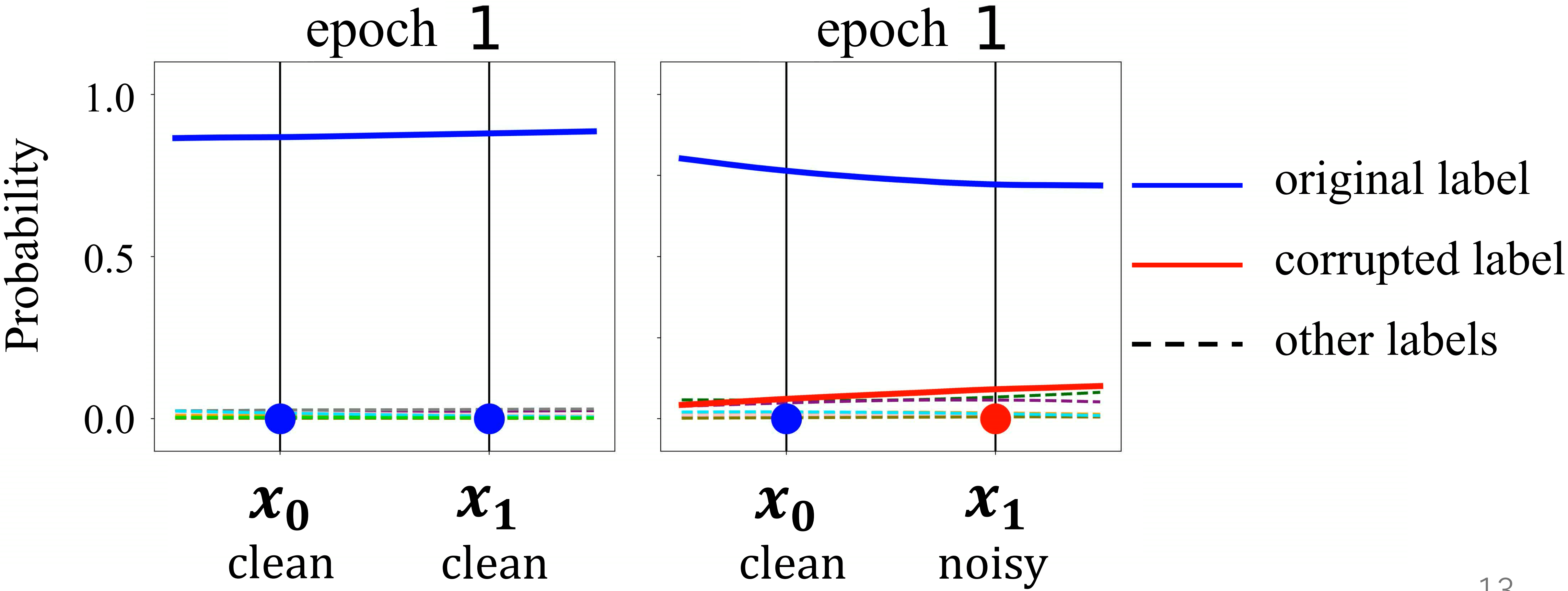


Why does performance on clean samples decrease and then recover?

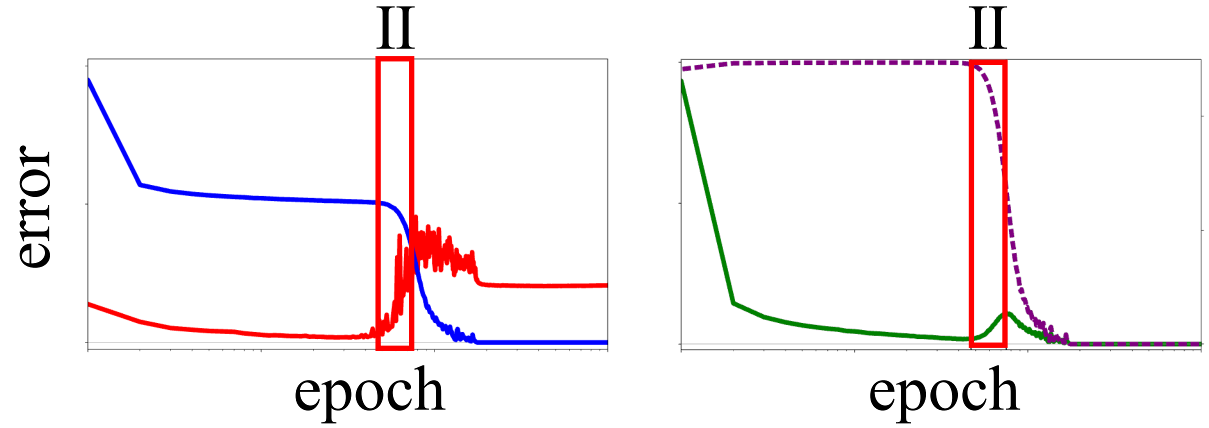
Phase I : Early training

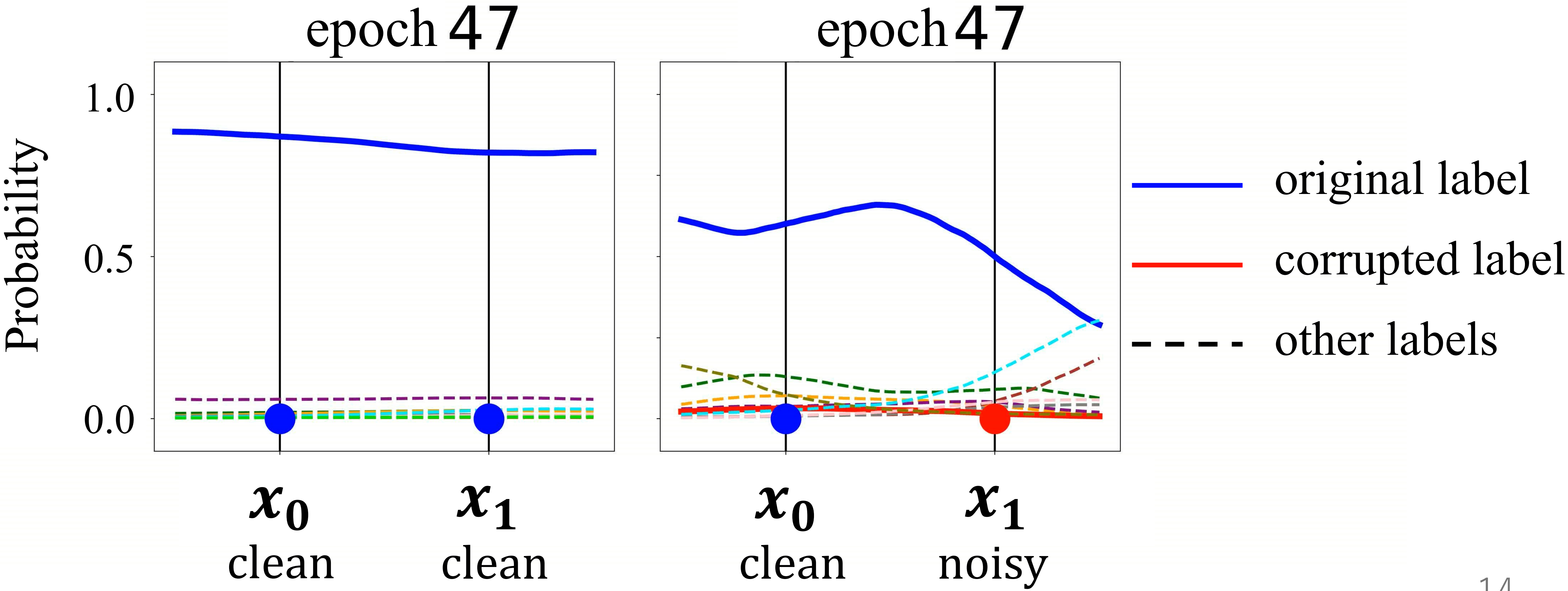




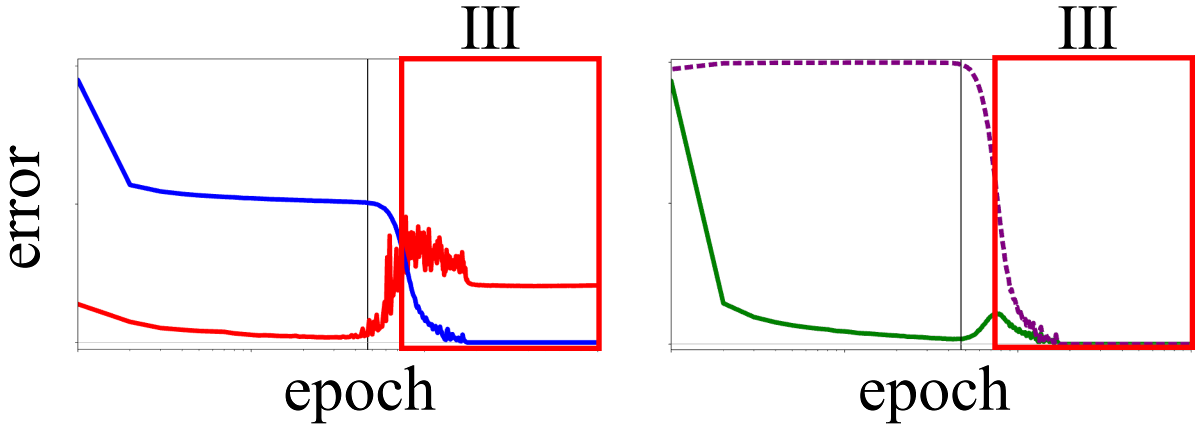


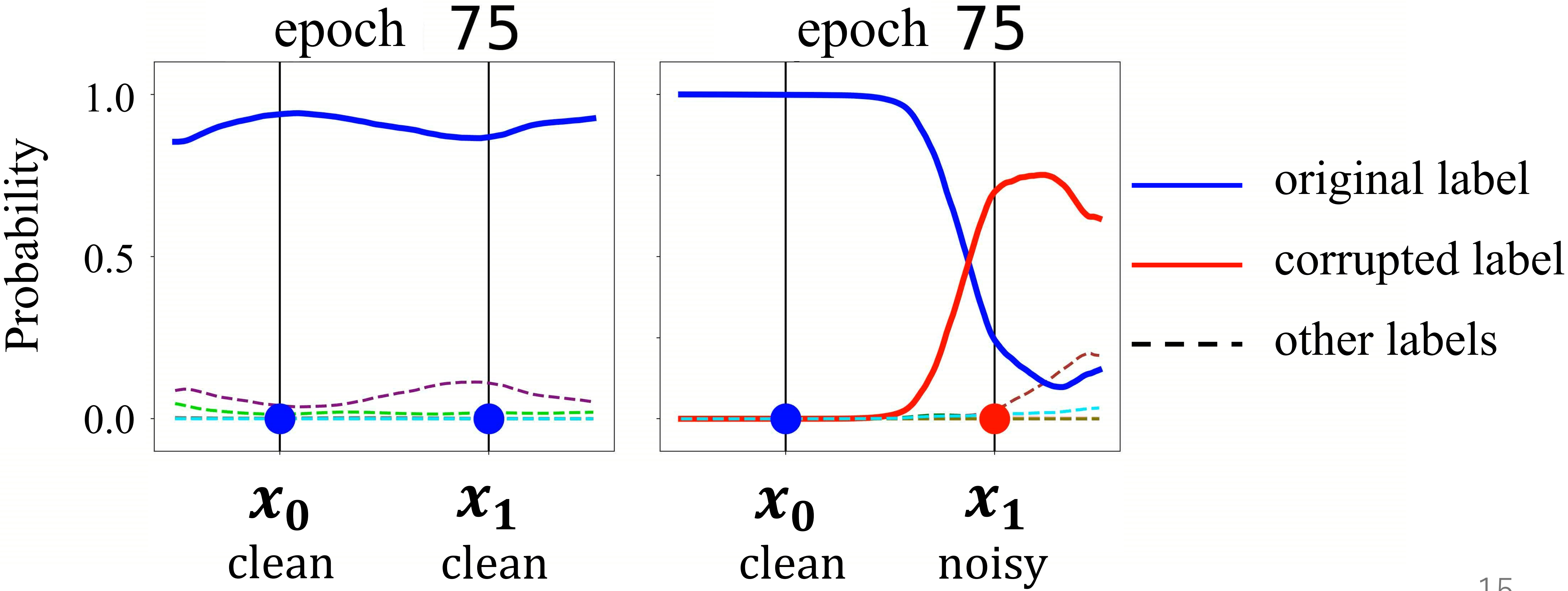
Phase II：Overfitting





Phase III：Double descent



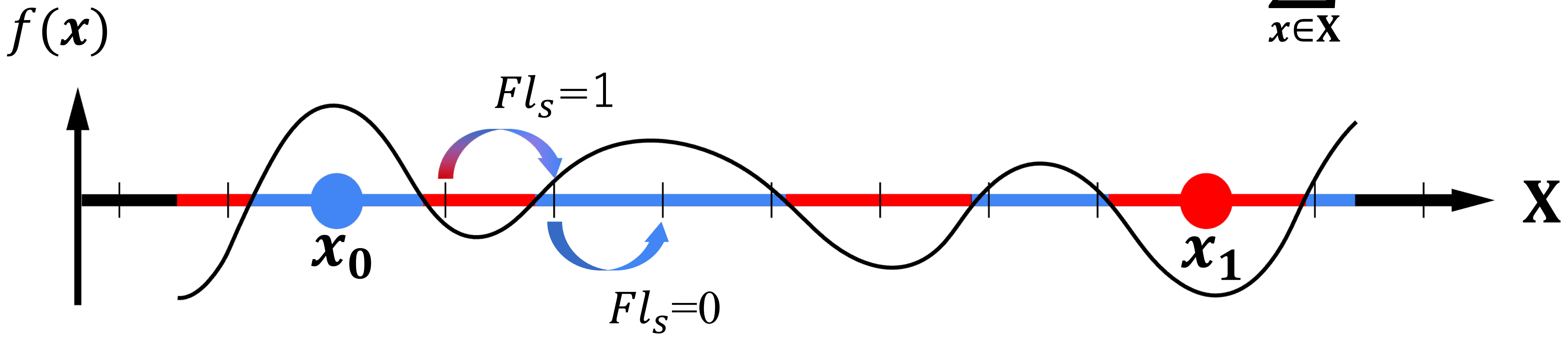


Spatial Instability : $INST_s$

Prediction variation within set **X** at time t

$$Fl_s(\mathrm{t}, \boldsymbol{x}) = \begin{cases} 1 & \text{if } \hat{f}_t(\boldsymbol{x}) \neq \hat{f}_t(\boldsymbol{x} - \Delta\boldsymbol{x}) \\ 0 & \text{if otherwise} \end{cases}$$

$$INST_s(\mathbf{X}, t) = \sum_{x \in \mathbf{X}} Fl_s(t, \boldsymbol{x})$$

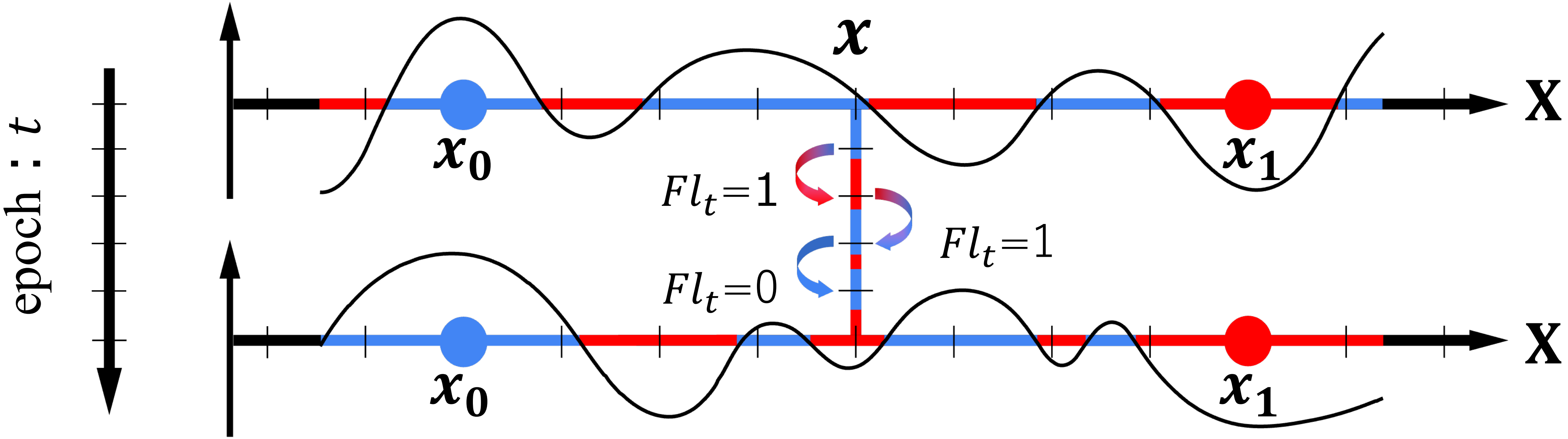


Temporal Instability : $INST_{\mathrm{T}}$

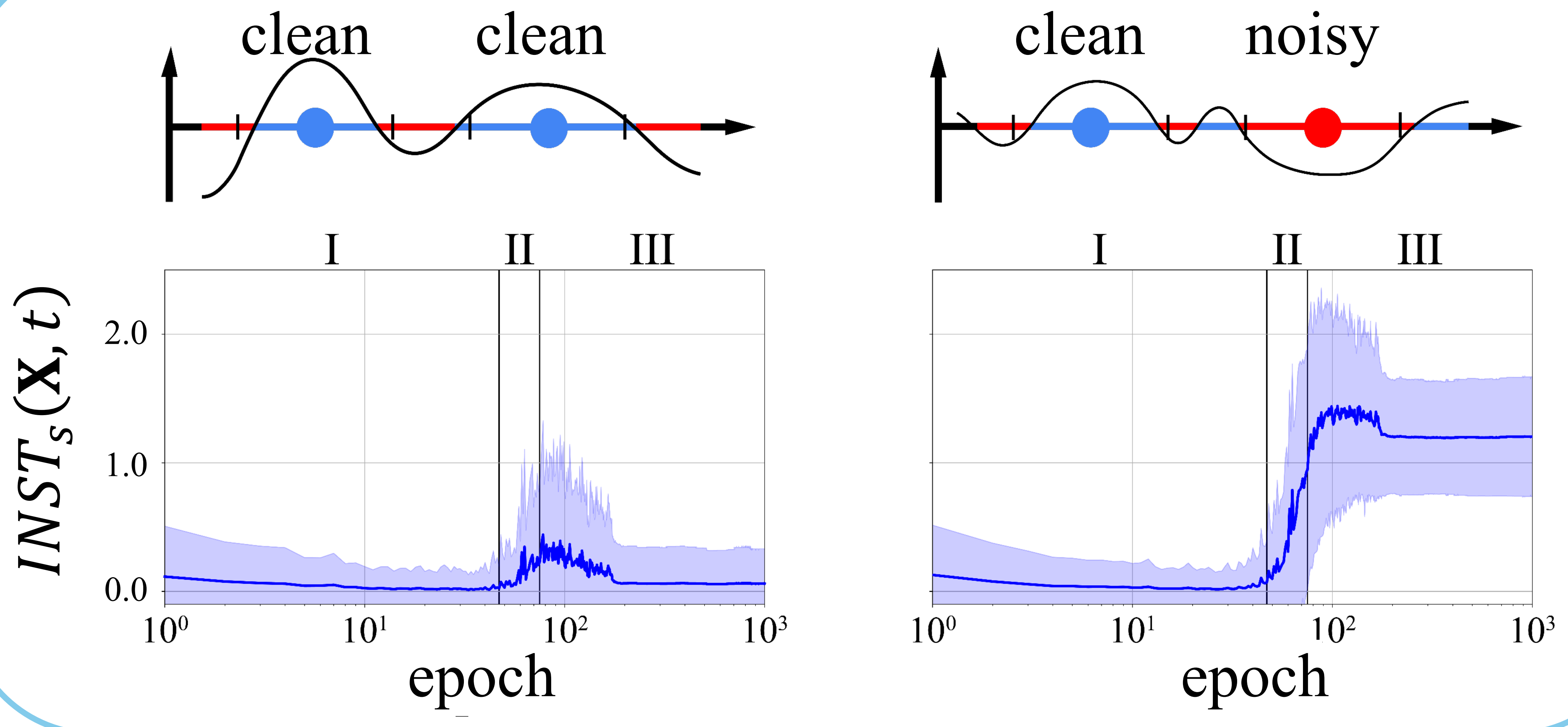
Prediction variation over time for data point $\boldsymbol{x}$

$$Fl_t(\mathrm{t},\mathrm{x}) = \begin{cases} 1 & \text{if } \hat{f}_t(\boldsymbol{x}) \neq \hat{f}_{t+1}(\boldsymbol{x}) \\ 0 & \text{if otherwise} \end{cases}$$

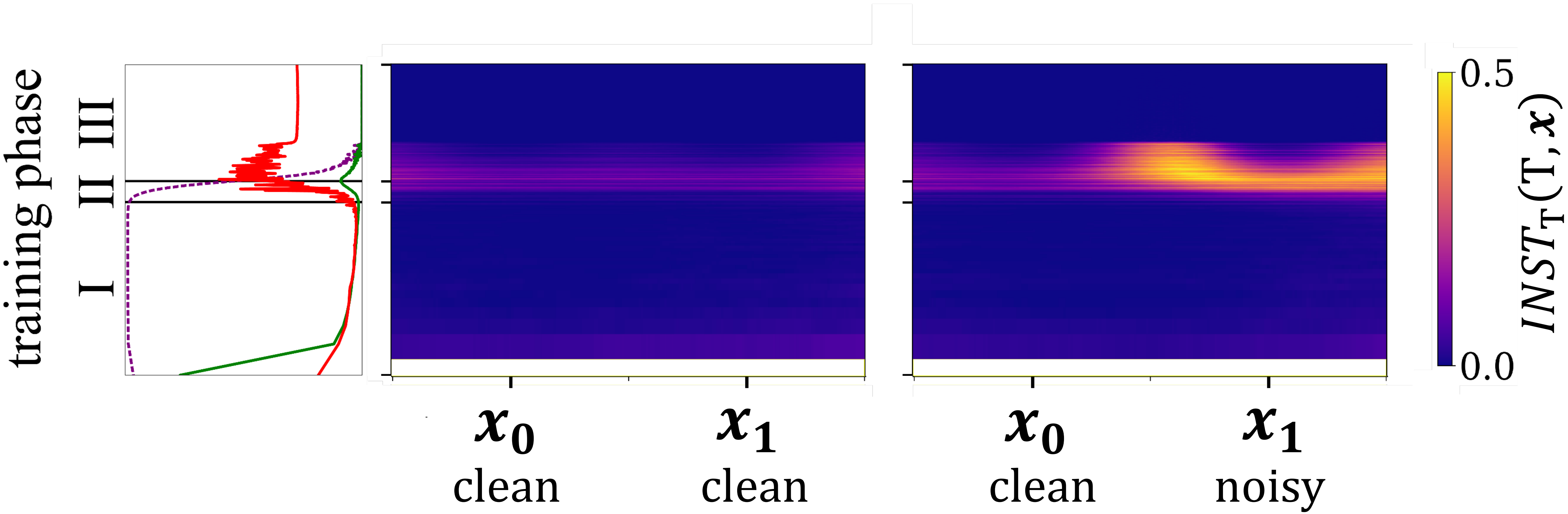
$$INST_{\mathrm{T}}(\mathrm{T},\boldsymbol{x}) = Fl_t(t,\boldsymbol{x})$$



Spatial Instability : $INST_s$



Temporal Instability： $INST_{\mathrm{T}}$



Discussion

- Clean/test performance **degradation** suggests that noisy-sample fitting causes **spatio-temporal fluctuations in the decision boundary.**

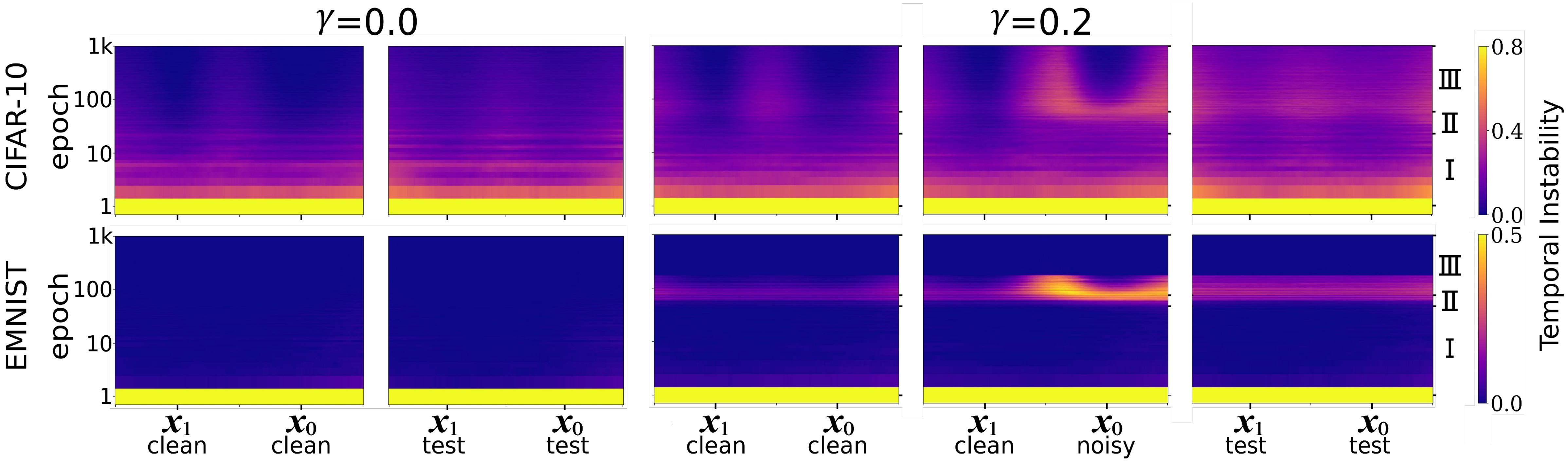
- We consider that the **stabilization** of these fluctuations leads to the **double descent** of the error rates on clean/test samples.

Conclusion

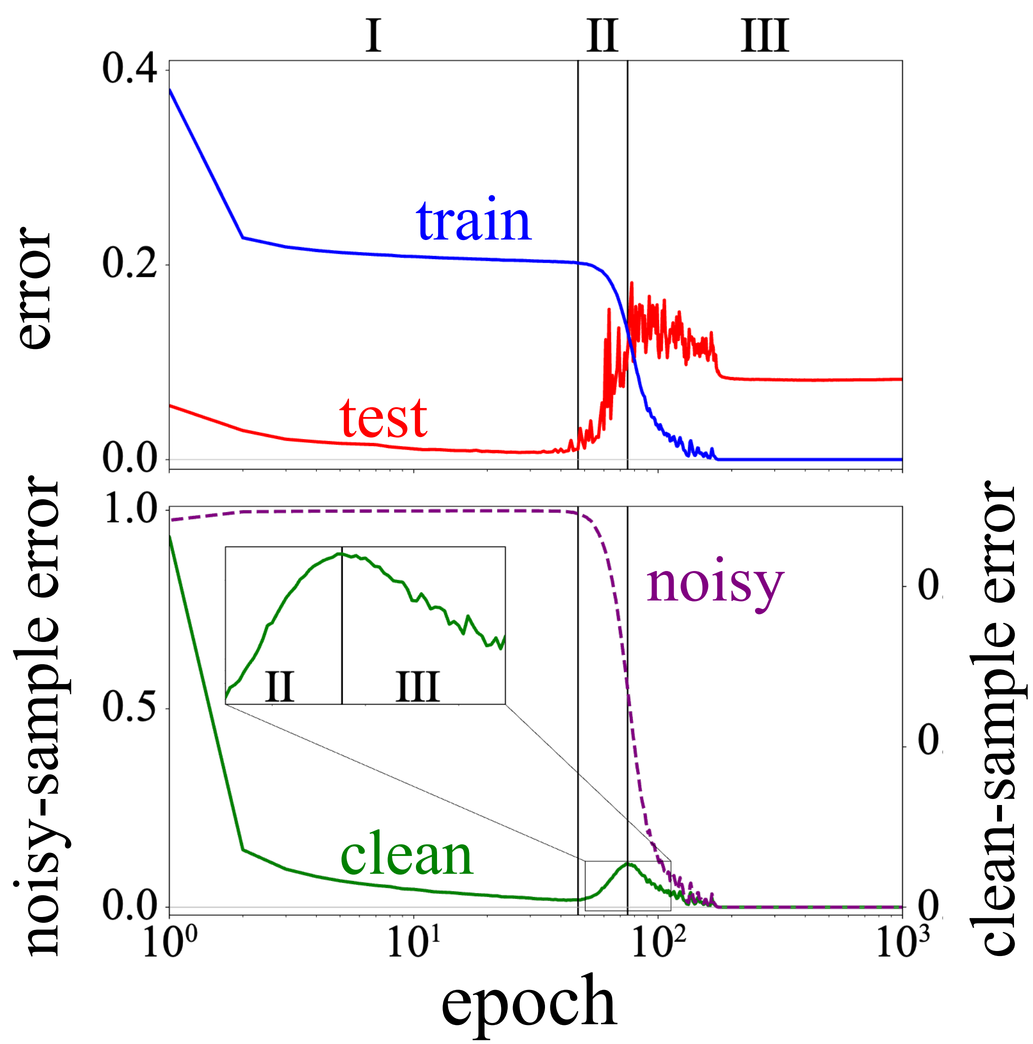
We showed that epoch-wise double descent occurs during the recovery of classification performance on clean samples.

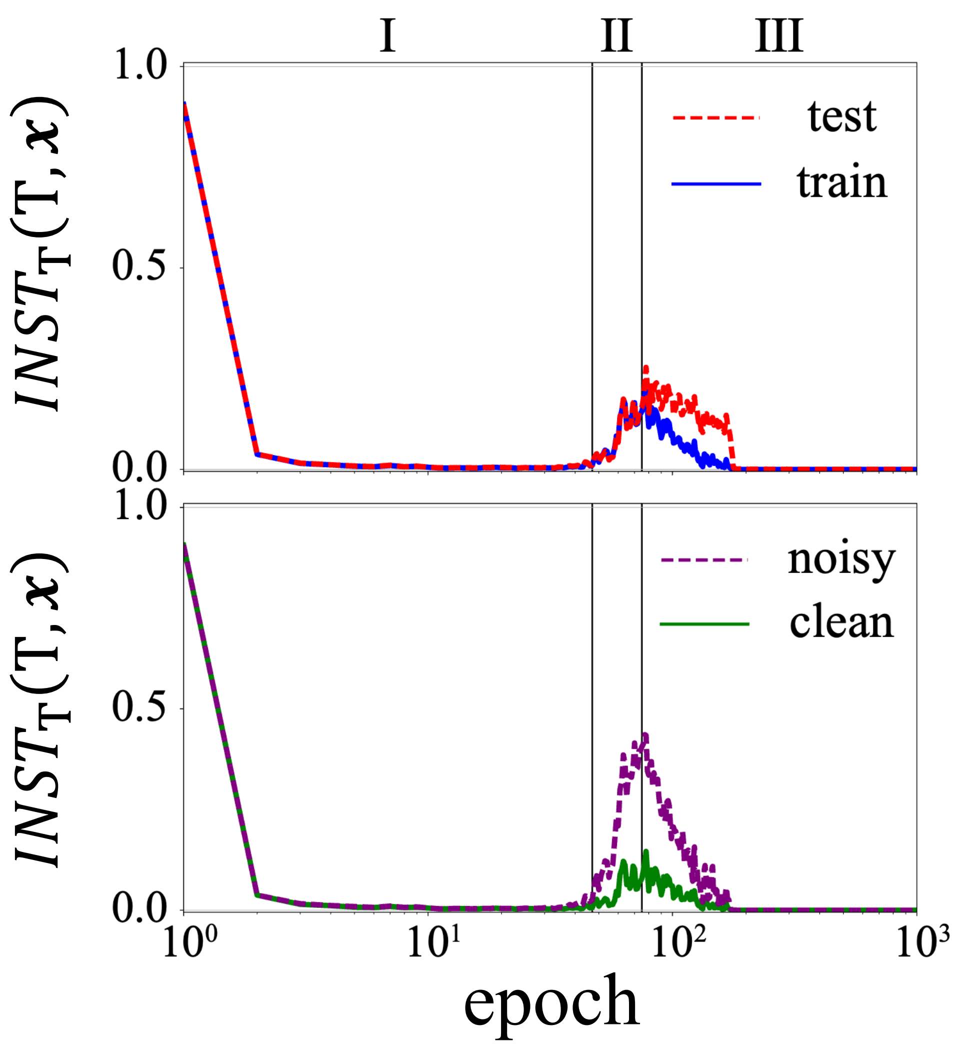
supplementary

Temporal Instability： $INST_{\mathrm{T}}$



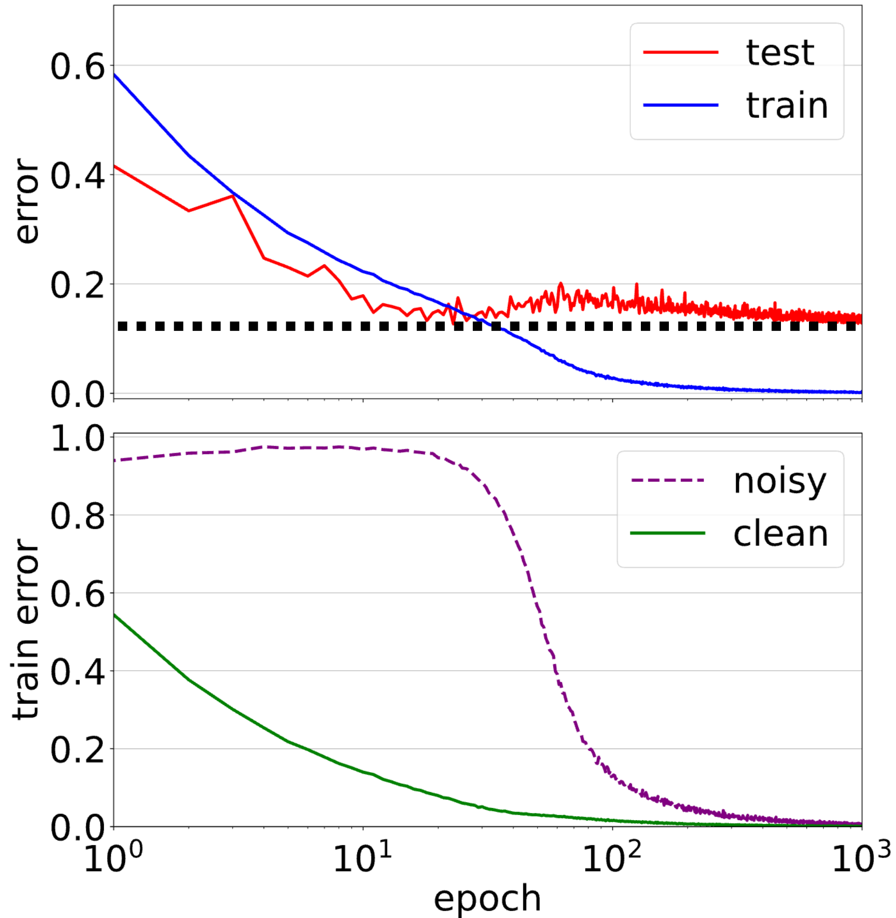
Temporal Instability : $INST_{\mathrm{T}}$





Performance after double descent

When the label noise rate is low, the test accuracy after double descent can be comparable to that before overfitting.



Model

ResNet-18

learning rate: 0.0001

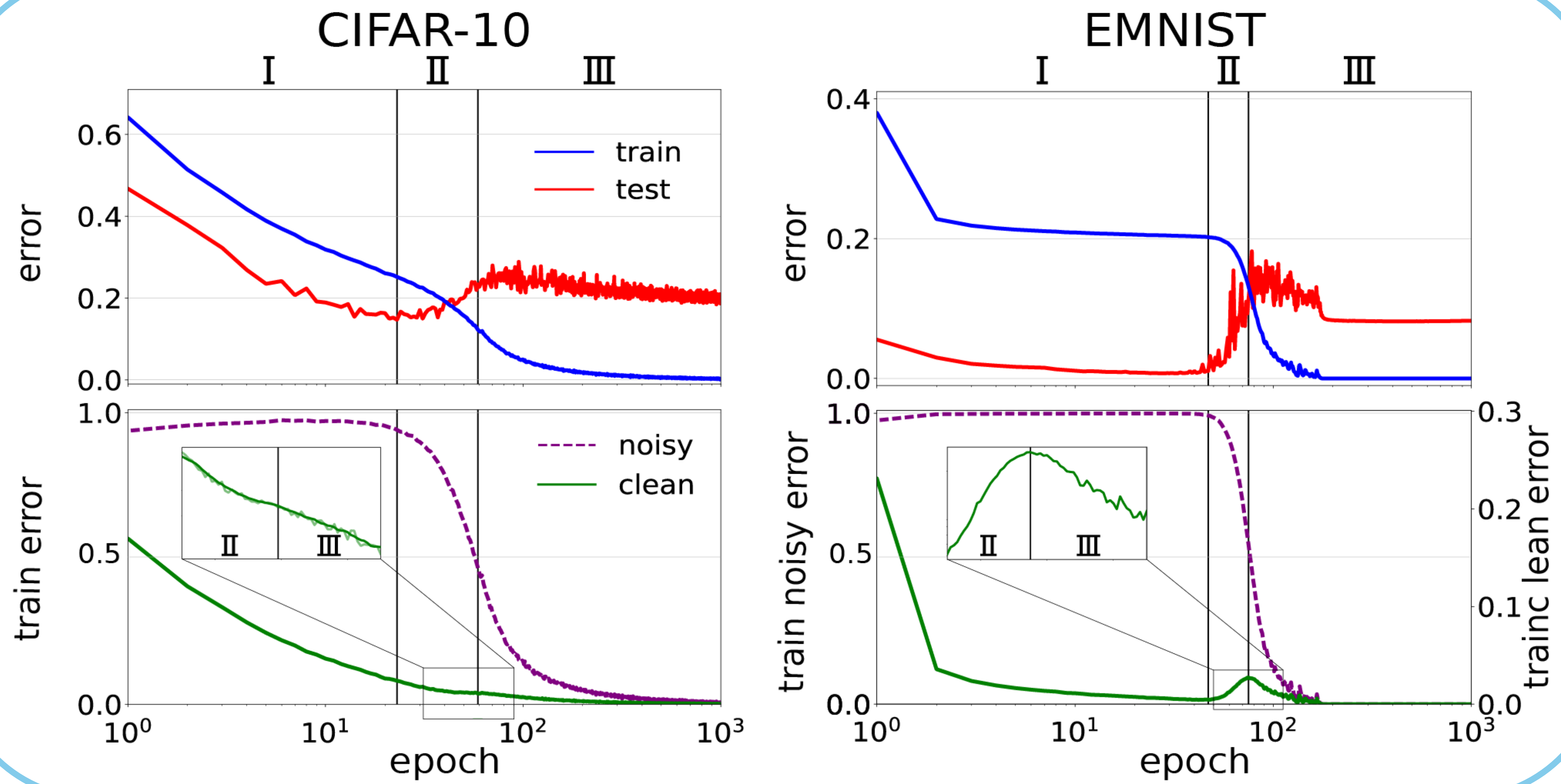
optimizer Adam

Dataset

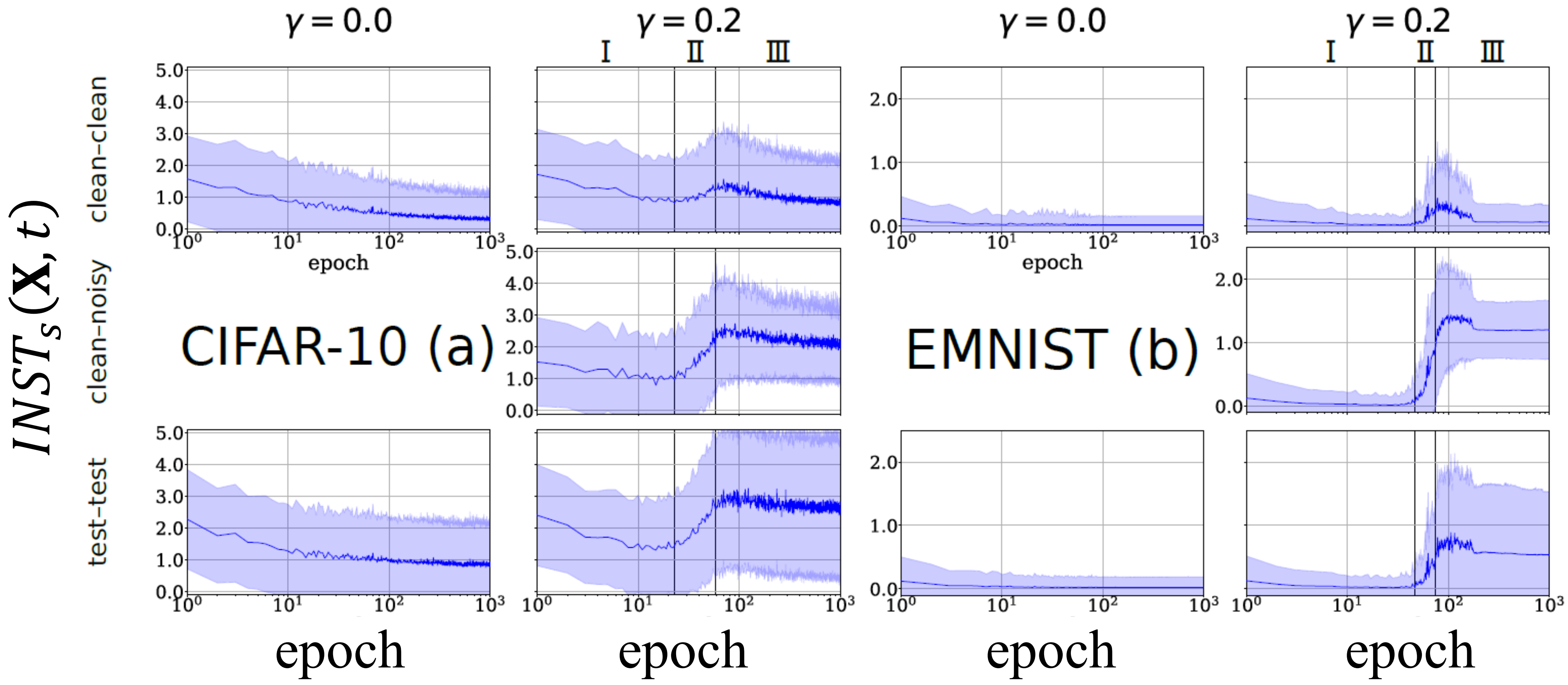
CIFAR-10

label noise rate 10%

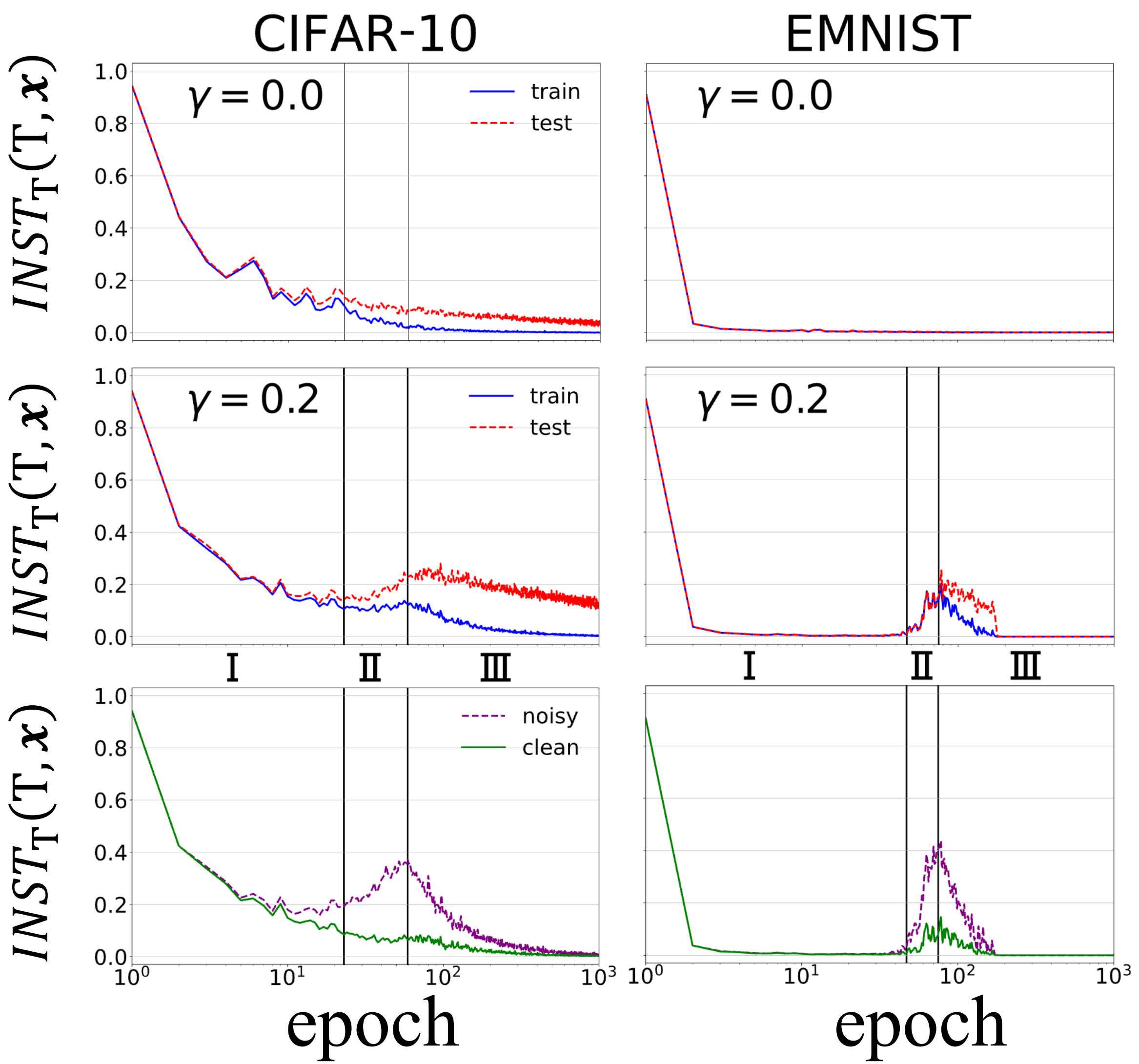
Nakkiran setting and Our Setting



Spatial Instability : $INST_s$

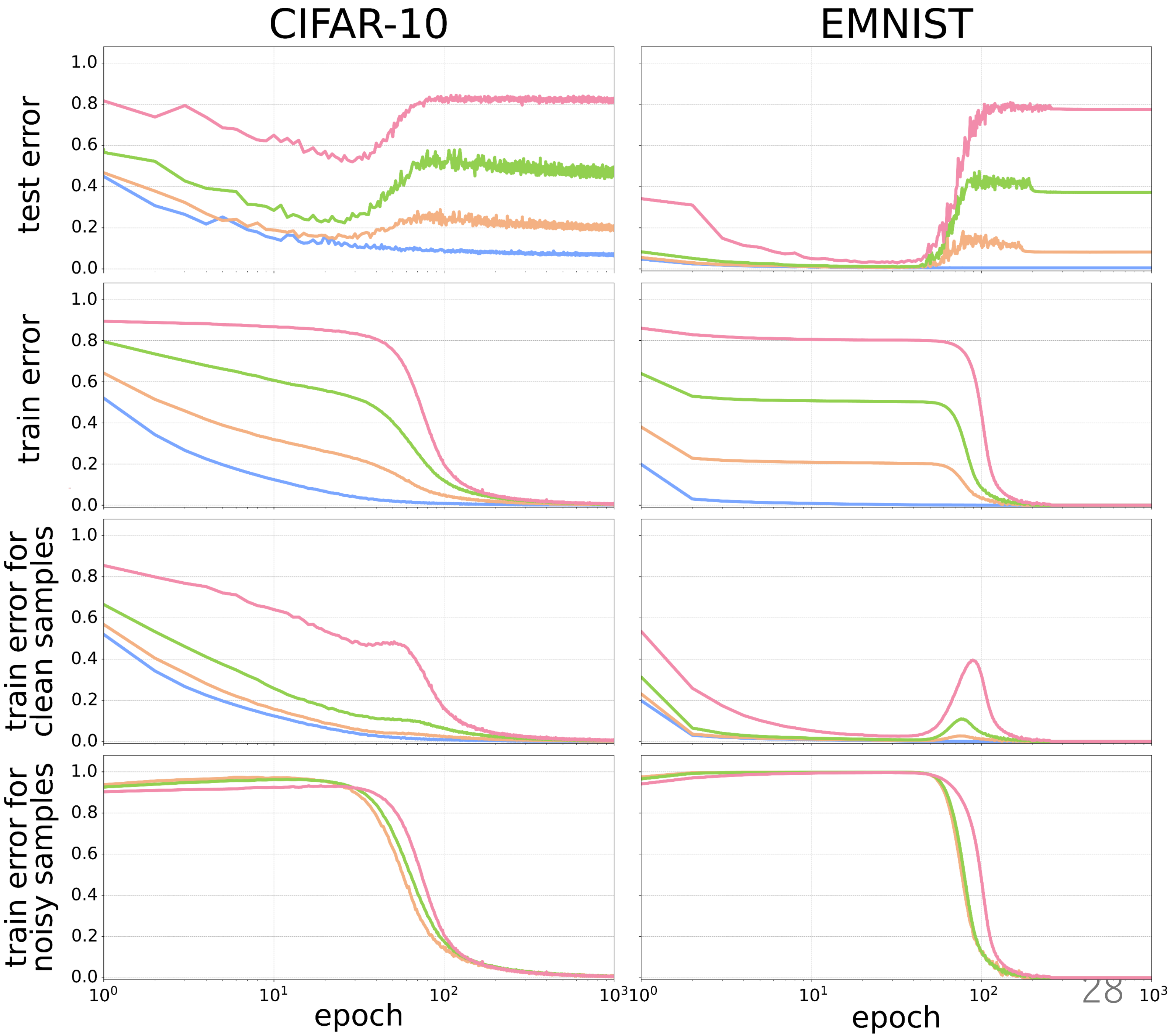


Temporal Instability : $INST_{\mathrm{T}}$

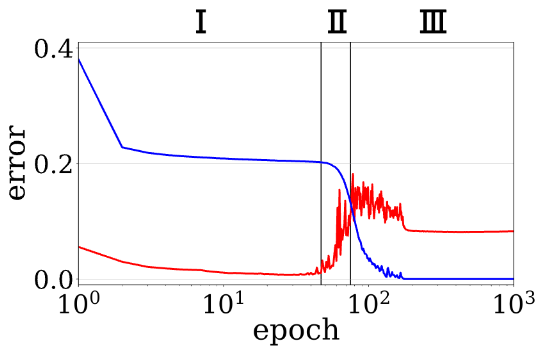


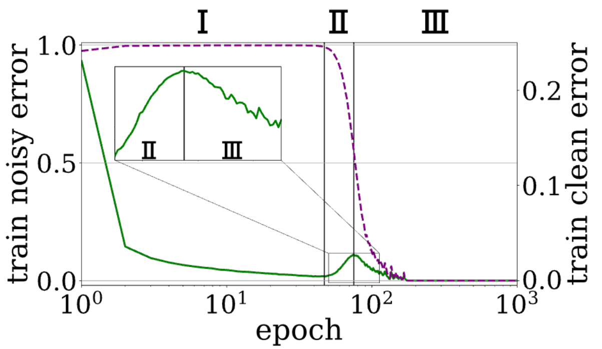
Various
Label noise rate

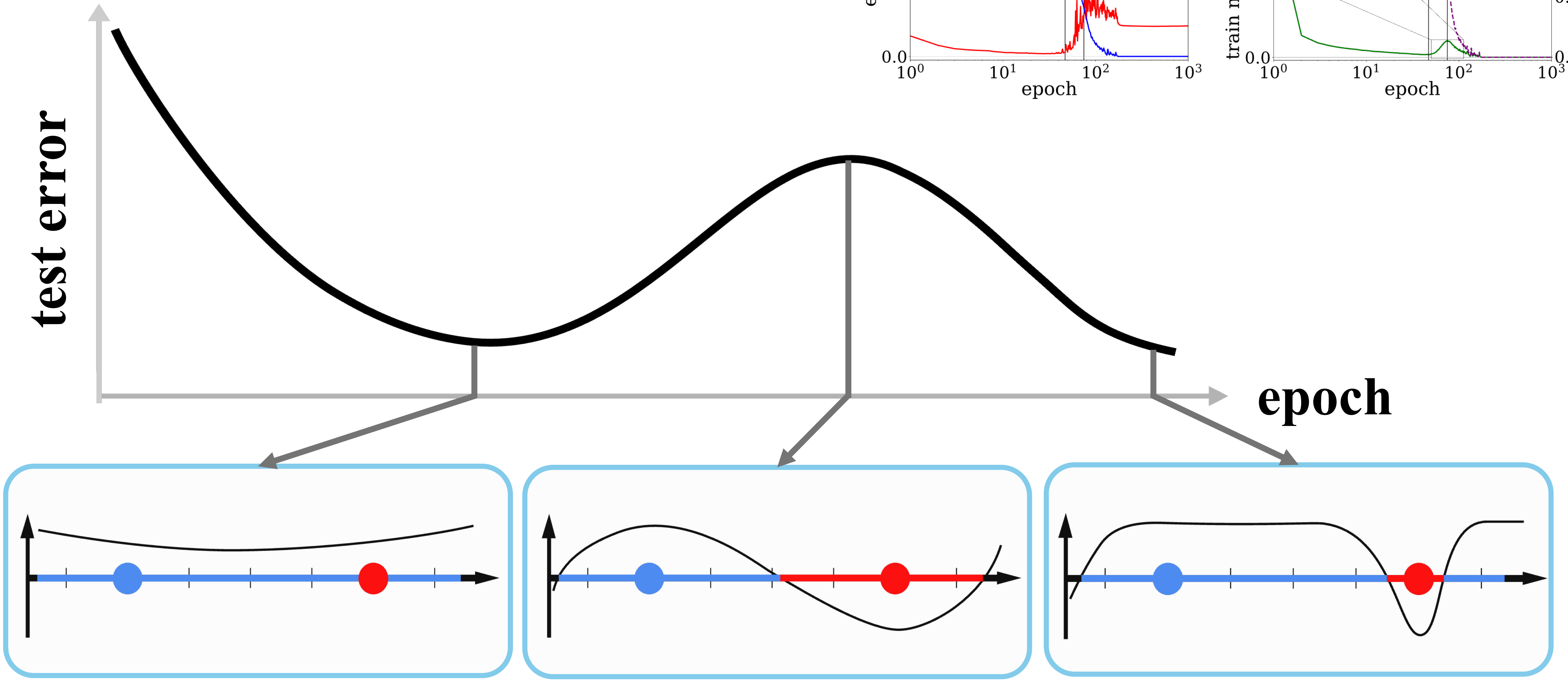
$\gamma = 0.8$
$\gamma = 0.5$
$\gamma = 0.2$
$\gamma = 0.0$



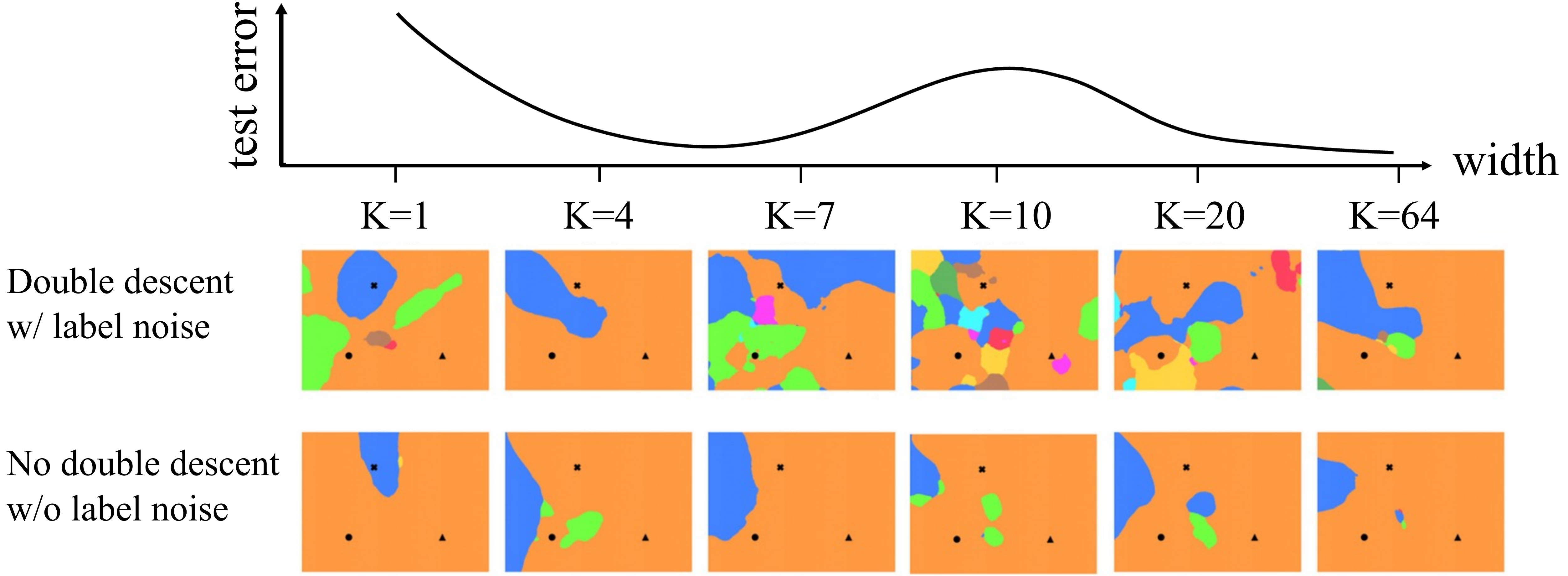
Initial hypothesis







Related Work[3]



[3] Somepalli+ （CVPR 2022）