* difficulties in DL generalization

- co-adaptation

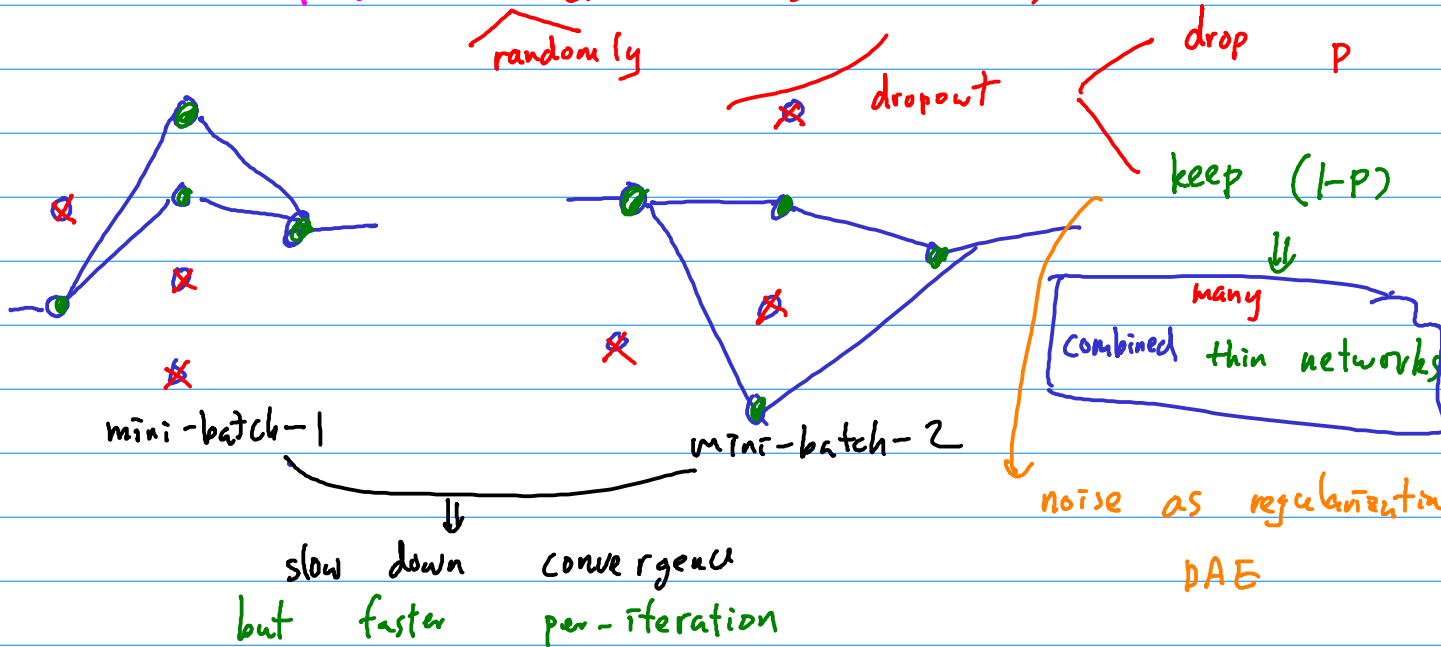(consistent mistakes from some neuron) correct by fitting other neurons

↓   ↓
noise   overfit
✗   ⌣ dependence   ✗

- break the dependence : shutdown some neurons
  ⌐randomly

dropout ✗   drop P
                   keep (1-P)

mini-batch-1   mini-batch-2   many
                                          combined thin networks

⇓   noise as regularization
slow down convergence   DAE
but faster per-iteration

* dropout testing

recorded
↓

$$E(s_i^{(\ell)}) = (1-p) s_i^{(\ell)}$$
in training   in testing

in-consistent

"pseudo"-dropout in testing

$$x_i^{(\ell)} = \theta(\boxed{(1-p)}s_i^{(\ell)})$$

inverted dropout

training : dropout $$x_i^{(\ell)} = \theta(s_i^{(\ell)}/(1-p))$$

testing : $$x_i^{(\ell)} = \theta((1-p)s_i^{(\ell)}/(1-p))$$
$$= \theta(s_i^{(\ell)})$$ "unchanged"