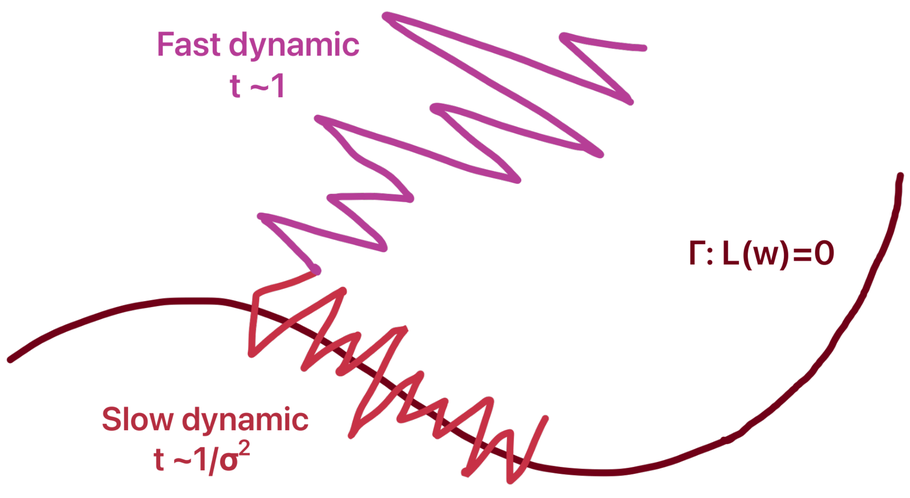
Dropout effectively adds
a regularizer to the loss.

Regularization Properties of Dropout

Anna Shalova, Mark Peletier and Andre Schlichting
a.shalova@tue.nl



Time-scale separation of the Dropout gradient descent near the zero-loss manifold.

Gaussian Dropout

Let $L : \mathbb{R}^k \to [0, \infty)$ be a differentiable loss functions.
Gaussian Dropout Gradient Descent is a modification of Gradient Descent:

$$w_{k+1} = w_k - \alpha \nabla L_w(w_k), \qquad w_{k+1} = w_k - \alpha \nabla \hat{L}_w(w_k, \eta_k), \quad \eta_{k,i} \sim \mathcal{N}(0, \sigma^2).$$

where $\hat{L} : \mathbb{R}^{k+d} \to [0, \infty)$ satisfies $\hat{L}(w, 0) = L(w)$.

Note: $\hat{L}$ is not unique.

Zero-loss manifold

Consider the gradient flow and correponding solution map:

$$\dot{w} = -\nabla L(w), \qquad \phi(w, t) = w - \int_0^t \nabla L(\phi(w, s)) ds.$$

Whenever exists introduce

$$\Phi(w) = \lim_{t \to \infty} \phi(w, t).$$

We assume that the zero-loss set $\Gamma = \{x : L(x) = 0\}$ is an M-dimensional C^2 manifold that satisfies some non-degeneracy assumptions and $\Phi(w)$ is well-defined in some neighboughood of Γ.

Examples: feedforward neural networks, generalized linear models, etc.

Main Result

Time-rescaled process

$$w_n(t) = w_{\lfloor \frac{t}{\alpha_n \sigma_n^2} \rfloor}$$

converges to a gradient flow on the zero-loss manifold. More formally:

Theorem 1. *Under some technical assumptions on $\hat{L}$, let $\alpha_n, \sigma_n, \alpha_n/\sigma_n^2 \to 0$, then there exist an open $U \in \mathbb{R}^k : \Gamma \in U$ and for all compact $K \subset U$ we have*

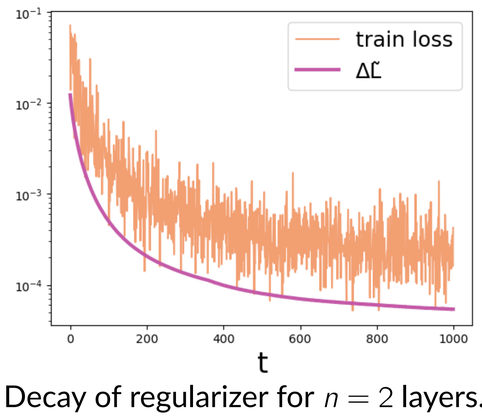
$$w_n^{\mu_n(K)} \Rightarrow w,$$

where $w(t) \in \Gamma$ and satisfies

$$w(t) = w(0) - \frac{1}{2} \int_0^{t \wedge \mu} P_\Gamma(w(s)) \nabla_w \Delta_\eta \hat{L}(w(s), 0) ds,$$

where $P_\Gamma(\xi)$ is an orthogonal projection onto the tangent space of Γ at $\xi \in \Gamma$.

Example



Decay of regularizer for $n = 2$ layers.

Consider n-layered ReLU neural networks with dropout filters η:

$$z^{k+1} = W^{k+1}(y^k \odot (1 + \eta^k)) + b^{k+1},$$
$$y^{k+1} = (z^{k+1})_+.$$

Then the regularizer for L_2 loss is:

$$\Delta_\eta \hat{L}(w, 0) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{n-1} \left\| \frac{\partial y^n}{\partial z^{k+1}} W^{k+1} y^k(x_i) \right\|^2.$$

For $n = 2$ the same result is derived in [ZX22].

Contributions:

- We generalize results of [ZX22] to a larger class of noisy systems.
- We prove convergence of the Dropout GD to the gradient flow restricted to the manifold of global minimizers.
- We also introduce Ornstein-Uhlenbeck dropout and study the effect of the noise-correlation on the performance of noisy GD.

References

[LWA21] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021.

[ZX22] Zhongwang Zhang and Zhi-Qin John Xu. Implicit regularization of dropout. *arXiv preprint arXiv:2207.05952*, 2022.

NWO
WWU MÜNSTER
TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY