Constraint-Based Regularization of Neural Networks

Benedict Leimkuhler, Timothée Pouchon, Tiffany Vlaar, Amos Storkey

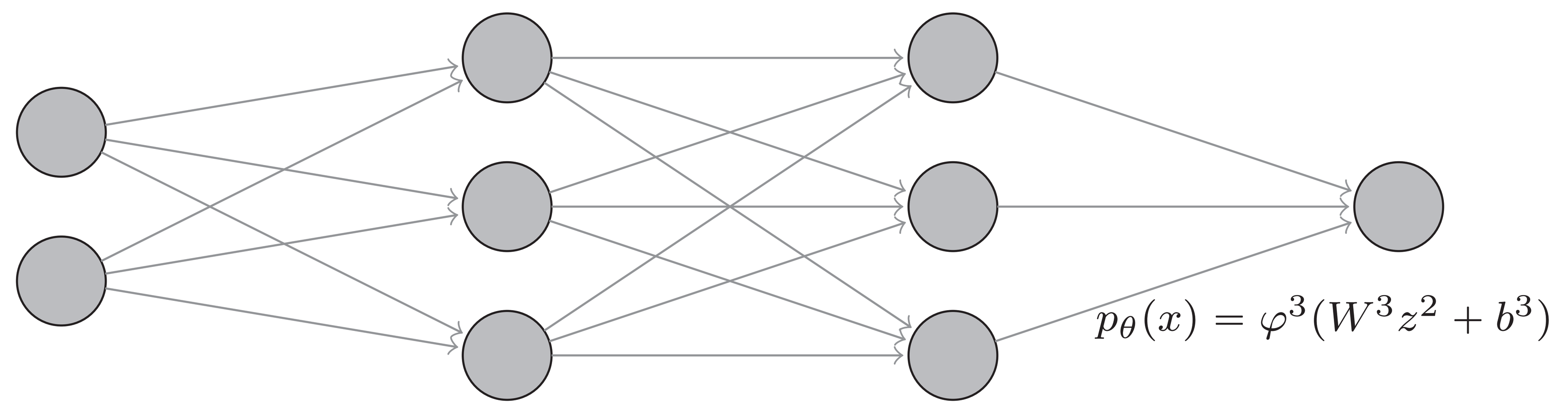
{b.leimkuhler, timothee.pouchon, tiffany.vlaar, a.storkey}@ed.ac.uk

Goal

Provide a mathematical framework for the regularization of deep neural networks by incorporating constraints into stochastic gradient Langevin dynamics.

Constraints for neural networks

Notations   Weights $W^\ell \in \mathbb{R}^{d^\ell \times d^{\ell-1}}$, biases $b^\ell \in \mathbb{R}^{d^\ell}$ $1 \leqslant \ell \leqslant L$



$z^0 = x$ $z^1 = \varphi^1(W^1 z^0 + b^1)$ $z^2 = \varphi^2(W^2 z^1 + b^2)$

Parameter vector: $\theta^\ell = (\mathrm{vect}(W^\ell), b^\ell) \in \mathbb{R}^{n^\ell}$, $\theta = (\theta^1, \dots, \theta^L) \in \mathbb{R}^{|n|}$, $|n| = \sum n^\ell$;

Slack variables $\xi \in \mathbb{R}^{n^\xi}$;   Training variable $q = (\theta, \xi) \in \mathbb{R}^d$, $d = |n| + n^\xi$.

We verify that the **gradient of the loss function** $L_X(\theta)$ is proportional to

$$\nabla^T_{\theta^L} p_\theta(x) = F^L_x P^L_x, \quad \nabla^T_{\theta^\ell} p_\theta(x) = F^L_x W^L \cdots F^{\ell+1}_x W^{\ell+1} F^\ell_x P^\ell_x \quad \ell \leqslant L-1,$$

where F^j_x is a sparse matrix evaluated from $\nabla \varphi^j$, and P^j_x is also sparse.
To control the above expressions, we may impose constraints on the parameter space: for a field $g : \mathbb{R}^d \to \mathbb{R}^m$, the **constraint manifold** is defined as

$$\Sigma = \{q \in \mathbb{R}^d \mid g(q) = 0\}.$$

Circle constraints: restrict each constrained parameter as $|\theta^c_i| \leqslant r_i$, where $r_i > 0$ is a given hyperparameter:

$$g_i(q) = |\theta^c_i|^2 + |\xi_i|^2 - r_i^2 \qquad 1 \leqslant i \leqslant m.$$

Orthogonality constraints: for a specific layer ℓ, we define

$$g(q) = \begin{cases} (W^\ell)^T W^\ell - I_{d^{\ell-1}} & \text{if } d^{\ell-1} \leqslant d^\ell, \\ W^\ell (W^\ell)^T - I_{d^\ell} & \text{otherwise.} \end{cases}$$

Constrained SDEs

We define the potential $V(q) = L_X(\theta)$ and consider the **constrained overdamped Langevin** system

$$\begin{aligned} \mathrm{d}q_t &= -\nabla V(q_t)\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W_t - \nabla_q g(q_t)\,\mathrm{d}\lambda_t \\ 0 &= g(q_t) \end{aligned} \qquad \text{(CoLA-od)}$$

whose invariant measure is $\mathrm{d}\nu_\Sigma = Z^{-1} e^{-\beta V(q)}\,\mathrm{d}\sigma_\Sigma$, where σ_Σ is the surface measure of Σ and Z is the normalization constant.

Theorem (Exponential convergence to equilibrium). *Assume that there exists $\rho > 0$ such that*

$$\mathrm{Ric}_\mathfrak{g} + \beta \nabla^2_\mathfrak{g} V \geqslant \rho \mathfrak{g}. \quad (\mathfrak{g} \textit{ Riemannian metric}, \ \mathrm{Ric}_\mathfrak{g} \textit{ Ricci curvature}, \ \nabla^2_\mathfrak{g} V \textit{ Hessian})$$

Then there exists $R > 0$ such that $\langle \phi \rangle_{\nu_\Sigma} = \int_\Sigma \phi \,\mathrm{d}\nu_\Sigma$

$$\int_\Sigma \left| \mathbb{E}(\phi(q_t) \mid q_0) - \langle \phi \rangle_{\nu_\Sigma} \right|^2 \mathrm{d}\nu_\Sigma(q_0) \leqslant C(\phi) e^{-2R\beta^{-1} t} \qquad \forall \phi \in H^1(\nu_\Sigma),$$

where $C(\phi)$ depends only on ϕ.

An alternative to (CoLA-od) is the second order dynamics given by the **constrained underdamped Langevin** system:

$$\begin{aligned} \mathrm{d}q_t &= p_t\,\mathrm{d}t \\ \mathrm{d}p_t &= \big(-\nabla_q V(q_t) - \gamma p_t\big)\,\mathrm{d}t + \sqrt{2\gamma\beta^{-1}}\,\mathrm{d}W_t - \nabla_q g(q_t)\,\mathrm{d}\lambda_t \\ 0 &= g(q_t) \end{aligned} \qquad \text{(CoLA-ud)}$$

whose invariant measure is closely related to ν_Σ.

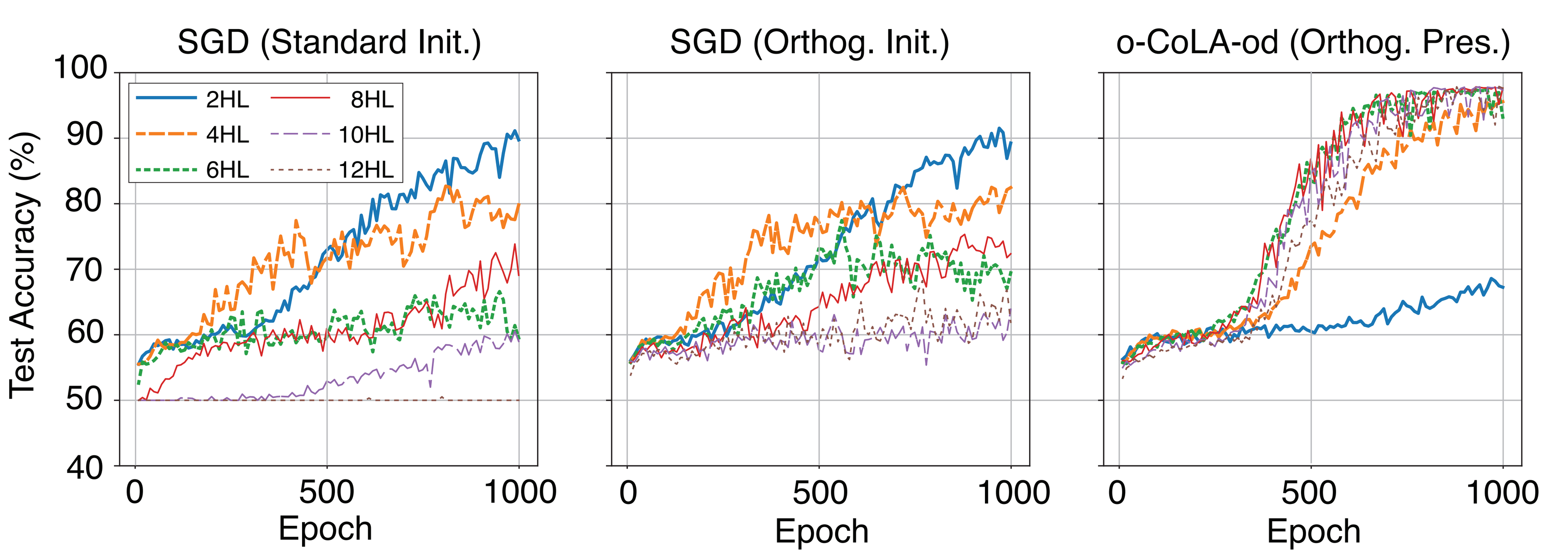
Example of discretization

Discretization of CoLA-od with orthogonality constraint (o-CoLA-od):

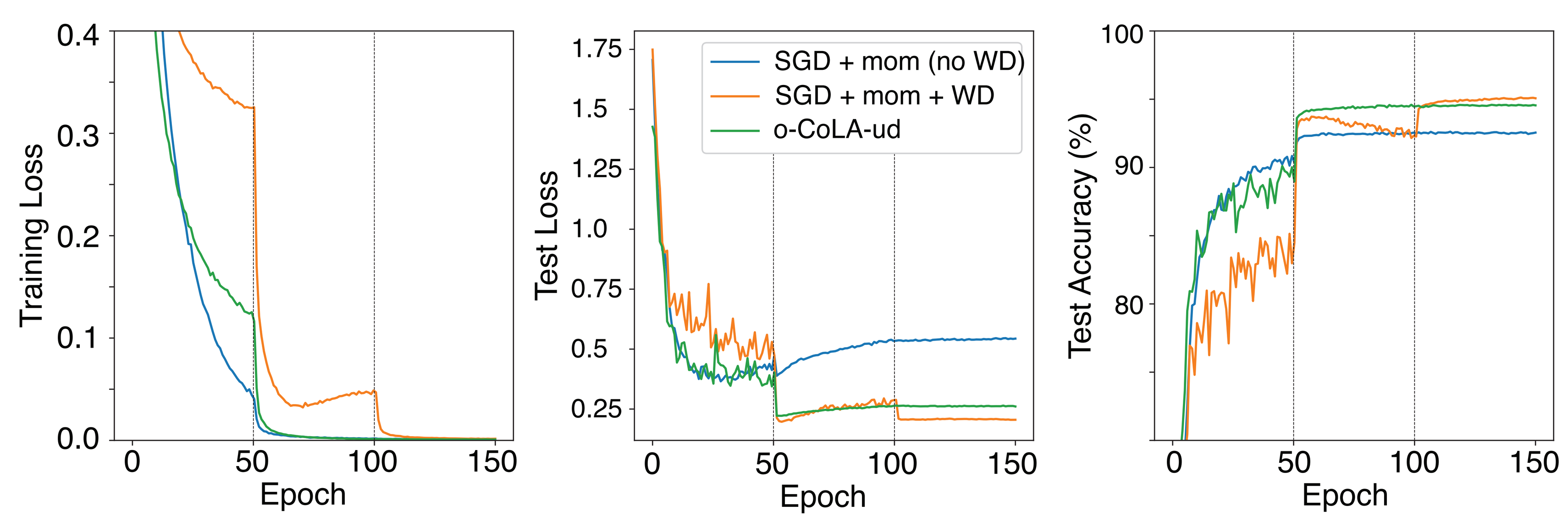
$$\begin{aligned} & Q^{(0)} = Q_n - h\nabla_Q V(Q_n) + \sqrt{2\tau h} R_n, \qquad (\tau = \beta^{-1}) \\ & \text{for } k = 0 \text{ to } K-1: \qquad Q^{(k+1)} = Q^{(k)} - \tfrac{1}{2} Q_n \big((Q^{(k)})^T Q^{(k)} - I_s \big), \\ & Q_{n+1} = Q^{(K)}. \end{aligned}$$

Numerical Experiments

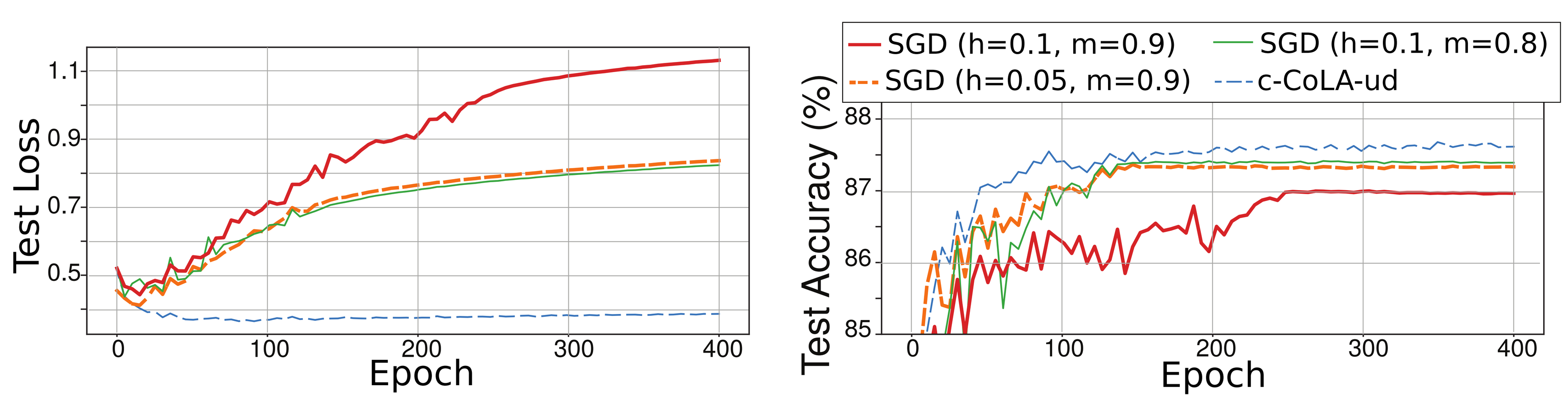
Spiral dataset: points in $\mathbb{R}^2$; 500 training points, 1000 test points.
Model: MLPs with variable depth, 100-nodes ReLU in each hidden layers.
Training: SGD with standard initialization (left), SGD with orthogonal initialization (middle) and o-CoLA-od with $\tau = 0$ (right).
Hyperparameters: for all methods, $h = 0.1$, 5% subsampling.
Results are averaged over 10 runs.



CIFAR-10 dataset: 32×32 images; 50K training images, 5K test images.
Model: ResNet-34 with BatchNorm.
Training: SGD-m vs. o-CoLA-ud with $\tau = 0$.
Hyperparameters: batchsize 128, lr-decay by a factor 10 every 50 epochs; for SGD-m, $h = 0.1$, mom. $= 0.9$; for o-CoLA-ud $\gamma = 0.5$, learning rate was rescaled to match the parameters of SGD-m.
Results are averaged over 5 runs.



Fashion-MNIST dataset: 28×28 images; # training images reduced to 10K, 60K test images.
Model: 1000-node SHLP.
Training: SGD-m vs. c-CoLA-ud with $\tau = 0$.
Hyperparameters: batchsize 128; for c-CoLA-ud, $h = 0.3$, $\tau = 0$, $\gamma = 1$, $r_0 = 0.05$, $r_1 = 0.1$.
Results are averaged over 5 runs.



Reference

[1] B. Leimkuhler, T. Pouchon, T. Vlaar, and A. Storkey, *Constraint-based regularization of neural networks.* arXiv:2006.10114, 2020.