Variational Learning Finds Flatter Solutions at the Edge of Stability

Avrajit Ghosh, Bai Cong, Rio Yokota, Saiprasad Ravishankar, Rongrong Wang, Molei Tao, Mohammad Emtiyaz Khan, Thomas Möllenhoff

University of California, Berkeley, RIKEN centre for Advanced Intelligent Project, Michigan State University, Institute of Science, Tokyo, Georgia Institute of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

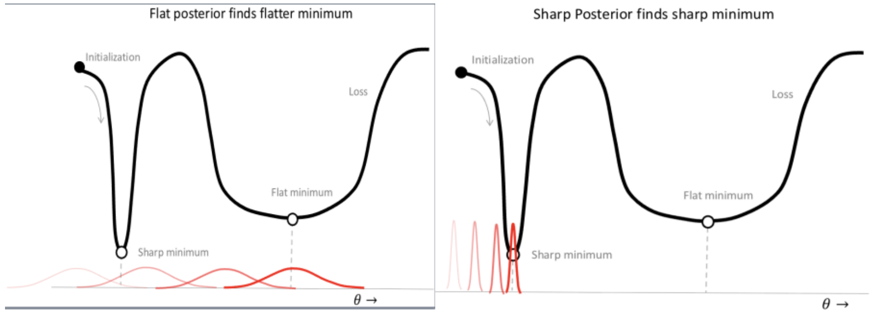
Contribution

- We study the training dynamics of Variational Learning and compare with Gradient Descent.

GD: $\theta_{t+1}=\theta_t-\rho\nabla\ell(\theta_t)$

VGD: $\mathbf{m}^{\varepsilon}_{t+1}\leftarrow\mathbf{m}_t-\rho\frac{1}{N_s}\sum_{i=1}^{N_s}\nabla\ell(\mathbf{m}_t+\varepsilon_i),\quad \varepsilon_i\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})\,.$

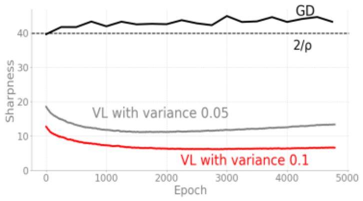
- VGD generalizes well due to its ability to find flat minima.
- How does VGD find flatter minima than GD? We understand the instability mechanism at play.
- We show the dynamical stability threshold for VGD is strictly smaller than GD, allowing it to find flatter minima.

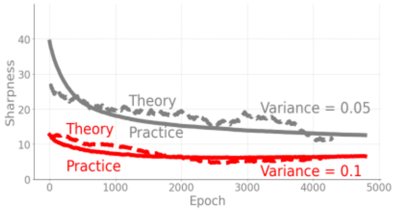


Variational Learning Mechanism of finding flatter solution

Instability dynamics in Deep Neural Network

- Gradient descent operates at "Edge of Stability" where loss decreases in oscillatory fashion.
- Sharpness of loss $\|\nabla^2\ell(\mathbf{m}_t)\|_2$ hovers about the threshold $2/\rho$.





- VGD operates at a smaller threshold than GD with increasing variance.

Descent on a Quadratic loss

$$\ell(\boldsymbol{\theta})=\frac{1}{2}\boldsymbol{\theta}^\top\mathbf{Q}\boldsymbol{\theta},\quad \text{where}\quad \mathbf{Q}=\sum_{i=1}^{d}\lambda_i\mathbf{v}_i\mathbf{v}_i^T.$$

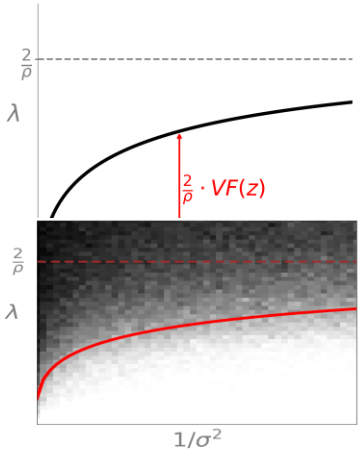
- GD decreases loss in one iteration if eigenvalues of Q are bounded by $2/\rho$
- VGD imposes regularization on the eigen-spectrum of the Hessian.

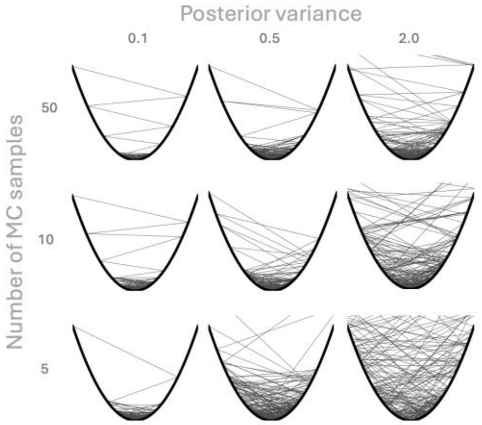
Descent condition for VGD

VGD update decreases loss quadratic loss

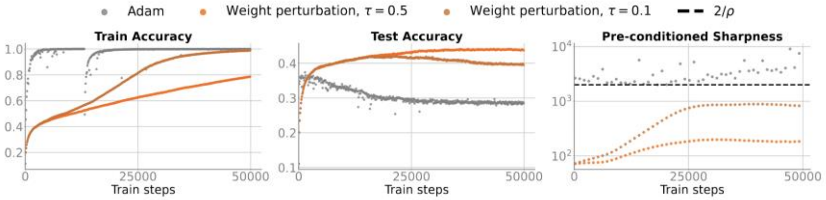
$$\mathbb{E}_{\varepsilon}[\ell(\mathbf{m}^{\varepsilon}_{t+1})]-\ell(\mathbf{m}_t)<0\quad\text{if}\quad\lambda_i<\frac{2}{\rho}\cdot\mathrm{VF}\left(\frac{N_s}{\sigma^2}\cdot c_{i,t}\right)\quad\text{for all } i,$$

With Variational Factor $\mathrm{VF}(z):=\rho\cdot\sqrt{\frac{z}{3}}\cdot\sinh\left(\frac{1}{3}\operatorname{arcsinh}\left(\frac{3}{\rho}\sqrt{\frac{3}{z}}\right)\right)$



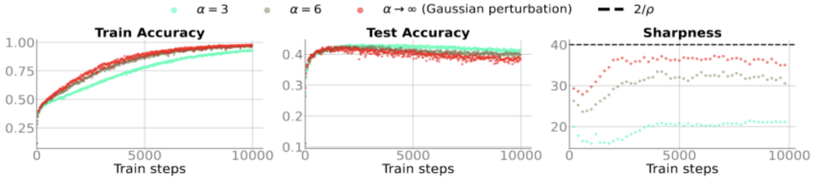


Adaptive Stability in Variational Online Newton

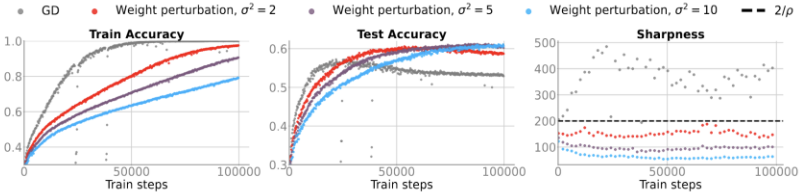


- Variational Online Newton learns distribution $q_t=\mathcal{N}(\mathbf{m}_t,\mathbf{P}_t^{-1})$ using Natural Gradient Descent.
- Adaptive Sharpness $\|\mathrm{diag}(\mathbf{p}_t)^{-1}\nabla^2\ell(\mathbf{m}_t)\|_2$ is smaller than $2/\rho$.

Heavy-tailed Student-t distribution



Stabilizing Attention Layers in Transformer


GD
Weight perturbation, $\sigma^2=2$
Weight perturbation, $\sigma^2=5$
Weight perturbation, $\sigma^2=10$
$2/\rho$
Train Accuracy
Train steps
Test Accuracy
Train steps
Sharpness
Train steps