Gradient Regularization Improves Accuracy of Discriminative Models

Dániel Varga, Adrián Csiszárik, Zsolt Zombori

Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest, Hungary

Highlights

- *Gradient regularization* means penalizing large gradients of the input-output mapping at the training data points.
- Our *SpectReg* gradient regularizer penalizes the squared Frobenius norm of the Jacobian of the classifier logits.
- SpectReg consistently improves generalization performance when the amount of training data is limited.
- On more complex tasks, it outperforms other gradient regularization variants such as
  - double backpropagation (Drucker and LeCun 1991), which penalizes the gradients of the loss function,
  - Jacobian regularization (Sokolic et al 2016), which penalizes the gradients of class probabilities.
- Spectral Regularization can be interpreted as "smart weight decay". Its effect is global, not confined to the neighborhood of the training points.

Background

Regularizing the gradient norm of a neural network's output with respect to its inputs is an old idea, going back to *Double Backpropagation* by Drucker and LeCun from 1991 (!). Variants of this core idea has been independently rediscovered several times since then, most recently by the authors of this paper. Most recent applications focus on robustness against adversarial sampling. Here we argue that gradient regularization can be used for the more fundamental task of increasing classification accuracy, especially when the available training set is small.
Here we present the two most promising variants: 1) classic *Double Backpropagation*, and 2) *Spectral Regularization* (SpectReg) which is our contribution.

Double Backpropagation (DoubleBack)

DoubleBack was discovered long ago, but its value as a regularizer on modern deep architectures has not yet been appreciated.
We take the original loss term and penalize the squared $L2$ norm of its gradient:

$$L_{DB}(x, y, \Theta) = L(x, y, \Theta) + \lambda \|(\frac{\partial}{\partial x} L(x, y, \Theta))\|_2^2$$

Although not immediately obvious from its definition, DoubleBack can be interpreted as applying a particular projection to the Jacobian of the logits and regularizing it.

Spectral Regularizer (SpectReg)

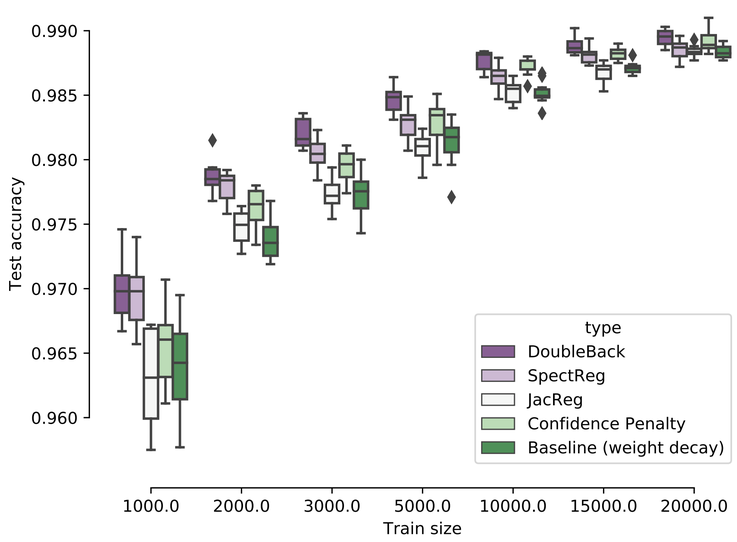
SpectReg is our own contribution. We found it to be the best variant for more complex datasets.
We apply a random projection to the Jacobian of the logits, and penalize the squared $L2$ norm of the result:

$$L_{SR}(x, y, \Theta) = L(x, y, \Theta) + \lambda \|P_{rnd}(J_g)\|_2^2$$

where $P_{rnd}(J_g) = J_g^T r$ and $r \in \mathcal{N}(0, I_m)$.
It's easy to show that SpectReg is an unbiased estimator of the squared Frobenius norm of the Jacobian.
The experiments we present control gradients at labeled training points, but we note that SpectReg makes no use of labels and can be applied to unlabeled inputs in a semi-supervised setting.

Gradient regularization and the size of the training set

The effect of regularizers is more significant for smaller training sets. However, as we demonstrate on MNIST the best gradient regularization variants (SpectReg and DoubleBack) maintain a significant benefit even for as much as 20000 training points.



Weight decay and gradient regularization

Gradient regularization yields significant benefit both in the presence and absence of weight decay, especially when the training set is small. The table below shows accuracy results restricted to 2000 training points.

Dataset	WD	Baseline	SpectReg
small MNIST	0	97.25 (0.22)	97.69 (0.11)
small CIFAR-10	0	48.27 (0.82)	50.41 (0.65)
small CIFAR-100	0	37.81 (0.33)	41.80 (0.70)
small MNIST	0.0005	97.40 (0.15)	**97.73** (0.13)
small CIFAR-10	0.003	55.63 (2.06)	**59.24** (1.48)
small CIFAR-100	0.003	49.56 (2.96)	**52.49** (0.65)

Gradient Regularization Compared with Dropout and Batch Normalization on small MNIST

Both DoubleBack and SpectReg achieve higher accuracy than either Dropout or Batchnorm in itself. We obtain the best result by combining Dropout and DoubleBack.

	NoGR	SpectReg	DoubleBack
Baseline	96.99 (0.15)	97.59 (0.13)	97.56 (0.24)
Batchnorm	96.89 (0.23)	96.94 (0.27)	96.89 (0.22)
Dropout	97.29 (0.19)	97.65 (0.14)	**97.98** (0.12)

Results on the TinyImageNet-200 dataset

Gradient regularization improves performance on the significantly more complex TinyImageNet-200 dataset. Note that in our various experiments the number of class labels range from 10 to 200, and the effect of gradient regularization does not fade with increased label count. This is noteworthy especially in the case of SpectReg, which is beneficial regardless of whether the number of matrix rows projected is 10 or 200.

Top-1 accuracy		Top-5 accuracy	
Baseline	SpectReg	Baseline	SpectReg
44.62 (0.58)	**50.76** (0.59)	70.20 (0.58)	**75.93** (0.40)

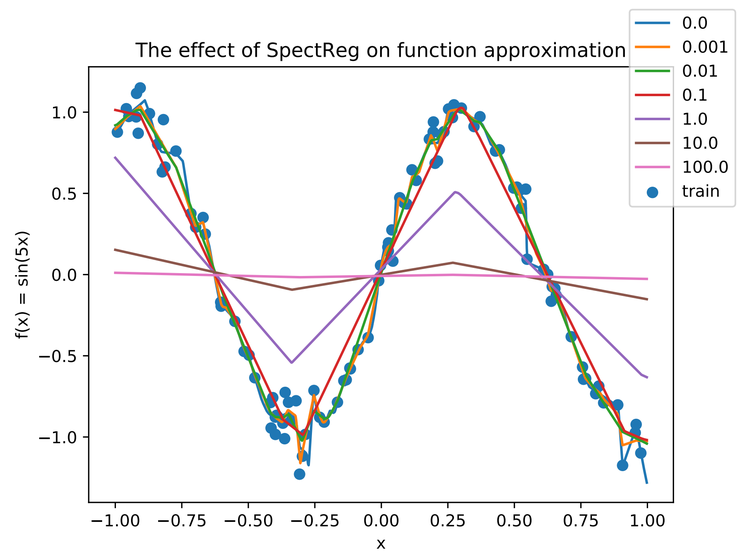
Gradient Regularization vs. Confidence Penalty

We compare Gradient Regularization and Confidence Penalty, a recently introduced regularization technique. All datasets are restricted to 2000 training points. On LeNet trained on MNIST, DoubleBack performs best. On a medium sized residual network trained on CIFAR-10/100, SpectReg clearly dominates.

Dataset	Baseline	SpectReg	DoubleBack	CP
MNIST	97.39	97.79	**97.89**	97.57
CIFAR-10	55.63	**59.24**	57.45	58.30
CIFAR-100	49.56	**52.49**	48.72	51.04

Local vs. global gradient control

A reasonable objection to gradient regularization methods is that they control the gradients only in the training points. In principle, a highly overparametrized network is capable of representing a "step function" that is extremely flat around the training points and contains unwanted sudden jumps elsewhere.
However, all our experiments indicate that gradient control smoothens the function on its whole domain. We emphasize that in all of our experiments except for TinyImagenet, the datasets are so small that the models are inevitably overfitting, yielding close to 100% train accuracy. In such setting, it is remarkable that gradient control acts globally. We demonstrate this here on a small synthetic dataset. Although the gradient is controlled only on 100 training points, the whole funtion becomes smoother. SpectReg reduces the mean squared error of the baseline model by a factor of 10.



SpectReg λ	0	0.001	0.003	0.01	0.03	0.1	0.3	1
MSE (1e-5)	16.6	24.5	2.5	2.3	**1.6**	3.4	74.9	497.7

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement 617747. The research was also supported by the MTA Rényi Institute Lendület Limits of Structures Research Group.
The project was supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

Contact Information

- daniel@renyi.hu
- csadrian@renyi.hu
- zombori@renyi.hu

Hungarian Academy of Sciences
Alfred Rényi Institute of Mathematics