Anticorrelated Noise Injection for Improved Generalization

Antonio Orvieto [*1] **Hans Kersting** [*2] **Frank Proske** [3] **Francis Bach** [2] **Aurelien Lucchi** [4]

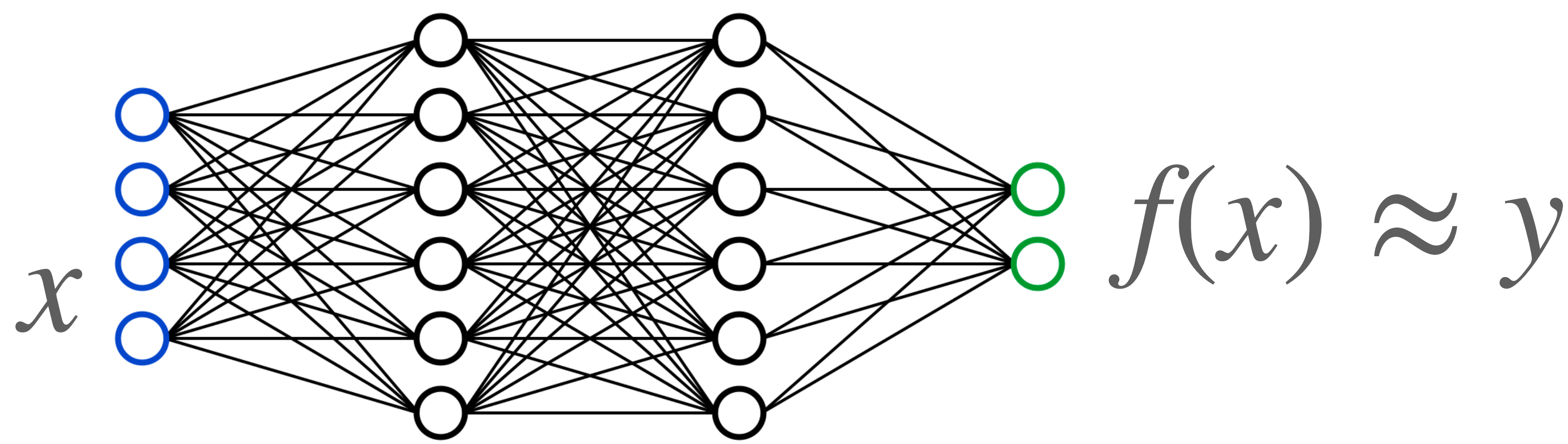




[*]Equal contribution [1]Department of Computer Science, ETH Zürich, Switzerland [2]INRIA - Ecole Normale Supérieure - PSL Research University, Paris, France [3]Department of Mathematics, University of Oslo, Norway [4]Department of Mathematics and Computer Science, University of Basel, Switzerland. Correspondence to: Antonio Orvieto <antonio.orvieto@inf.ethz.ch>, Hans Kersting <hans.kersting@inria.fr>.

Empirical Risk Minimization (ERM)

Let $f(x)$ be the prediction of a neural net which approximates the map $x \mapsto y$ for $(x, y) \sim P$



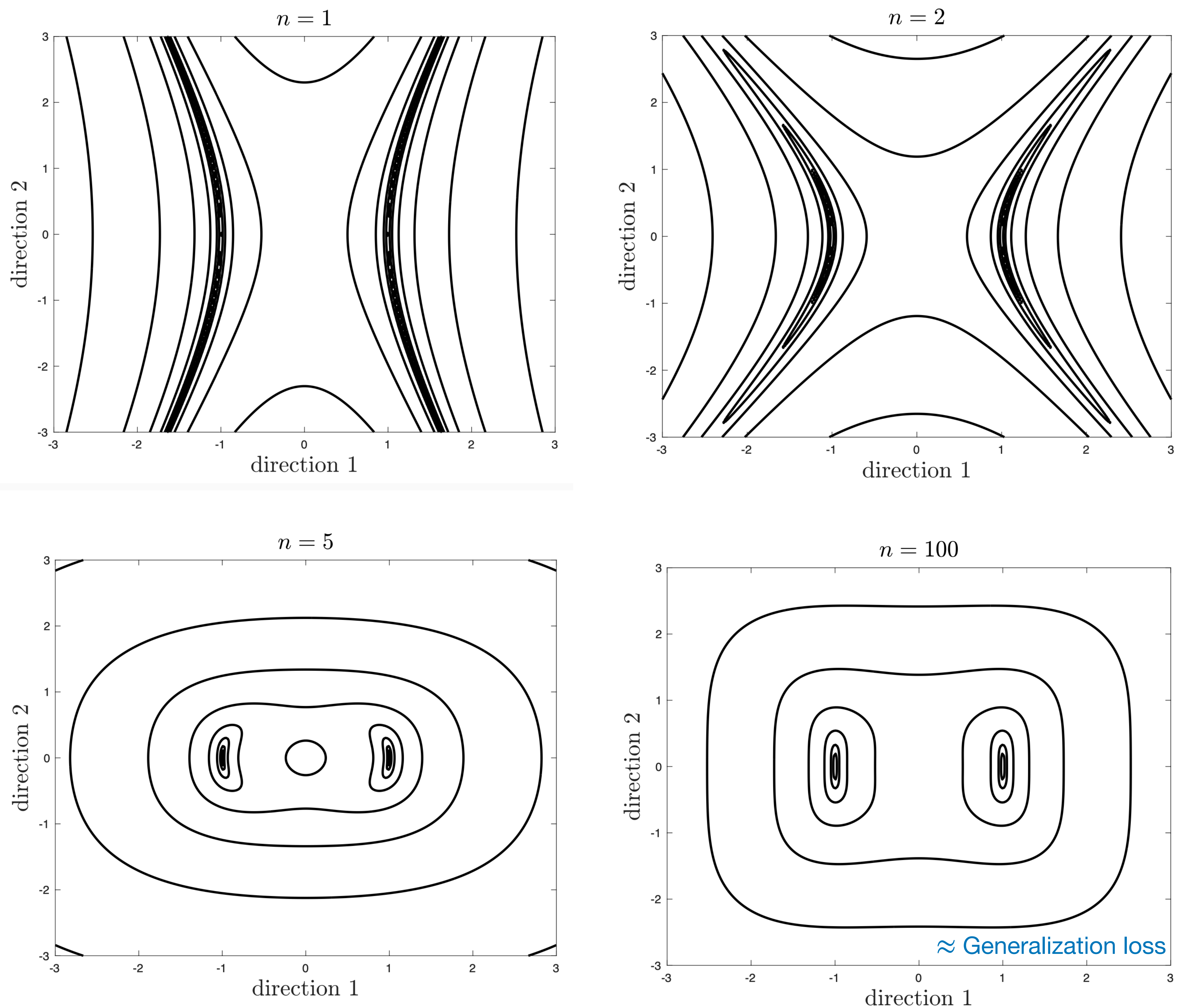
Consider a dataset $\{(x_i, y_i)\}_{i=1}^n$ sampled from P, we **hope** that

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} \ell_w(x_i, y_i).$$

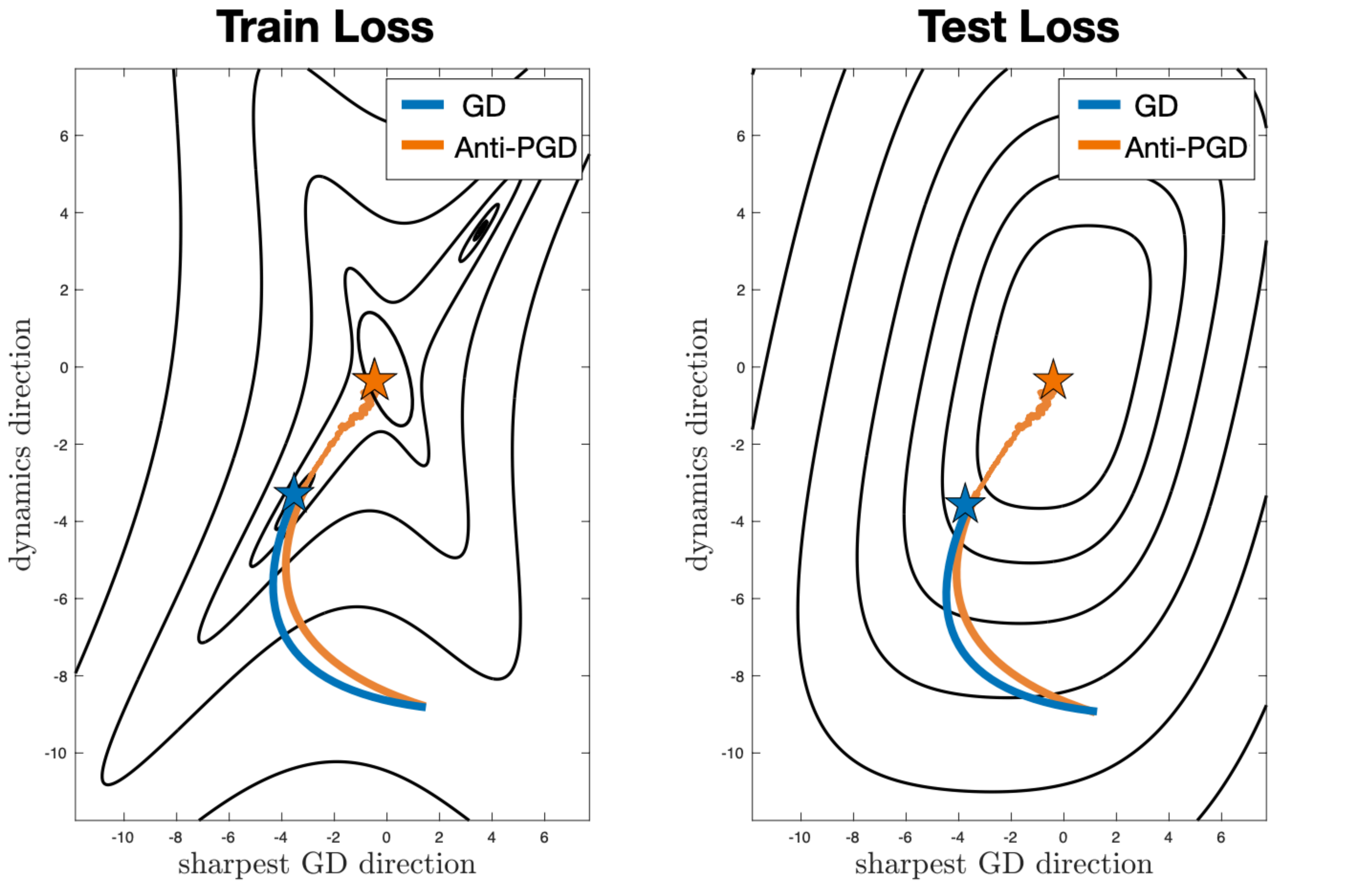
→ training loss

Is close to $L_{true}(w) = E_{(x,y)\sim P}[\ell_w(x, y)]$ → generalization loss

In over-parametrized models, loss landscape changes drastically as the number of datapoints increases!

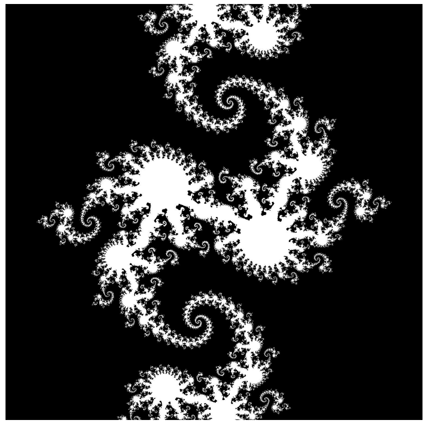


Proposed in this paper: Anti-PGD



Anti-PGD drives the approximation towards stable minima which provide improved generalization

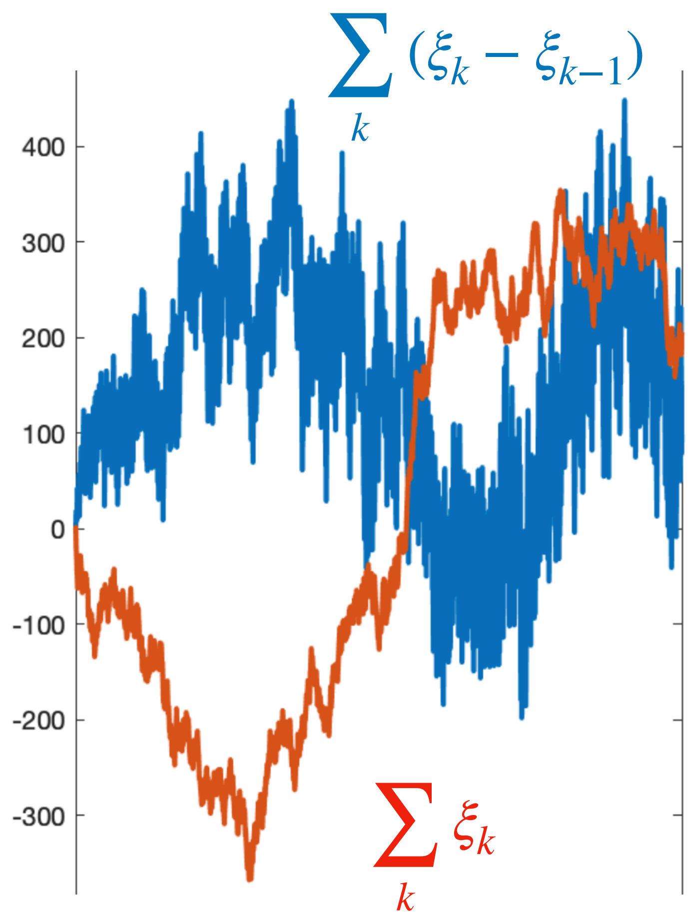
How are we able to do that?
Anti-correlated Noise Injection!



- Standard perturbed gradient descent (SGLD) is

$$w_{k+1} = w_k - \eta \nabla L(w_k) + \sigma \cdot \xi_{k+1}. \quad \textbf{(PGD)}$$

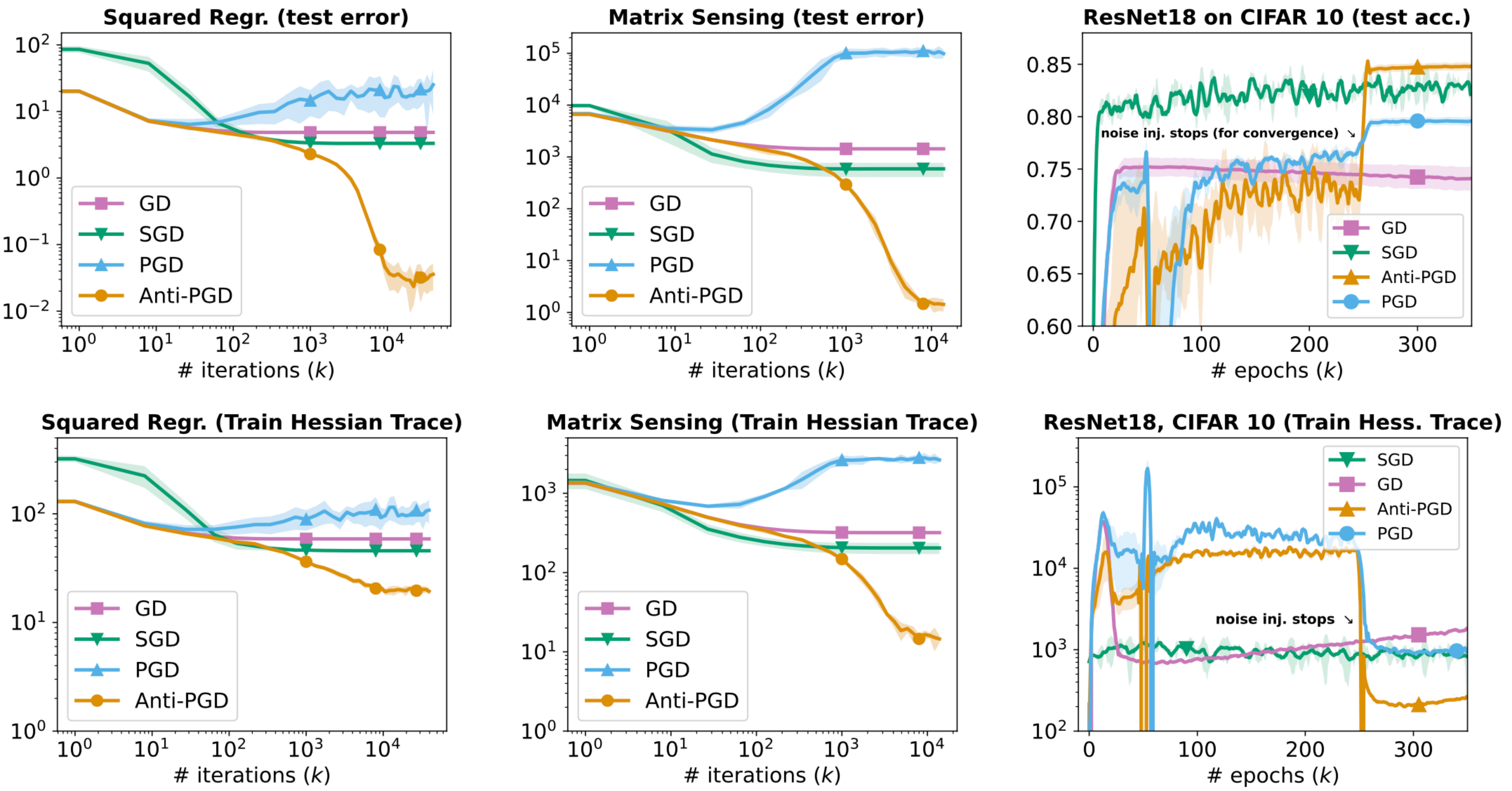
where ξ_k are standard Gaussian RVs.



- We negatively correlate noise to prev. update

$$w_{k+1} = w_k - \eta \nabla L(w_k) + \sigma \cdot (\xi_{k+1} - \xi_k) \quad \textbf{(Anti-PGD)}$$

Experimental evidence



Why does it work? (1)

Can be shown that adding anti-correlated noise corresponds to performing a noisy gradient step on a regularized loss

$$w_{k+1} = w_k - \eta \nabla L(w_k) + \sigma \cdot (\xi_{k+1} - \xi_k) \qquad \textbf{(Anti-PGD)}$$

$$\simeq w_k - \eta \nabla \tilde{L}(w_k) + \zeta_k, \qquad \zeta_k = \text{noise + h.o.t.}$$

Where $\tilde{L}$ is a regularized loss – penalises sharp minima!!

$$\tilde{L}(w) = L(w) + \frac{\sigma}{2} Tr(\nabla^2 L(w))$$

FLAT MINIMA

Neural Computation 9(1):1–42 (1997)

Sepp Hochreiter
Fakultät für Informatik
Technische Universität München
80290 München, Germany
hochreit@informatik.tu-muenchen.de
http://www7.informatik.tu-muenchen.de/~hochreit

Jürgen Schmidhuber
IDSIA
Corso Elvezia 36
6900 Lugano, Switzerland
juergen@idsia.ch
http://www.idsia.ch/~juergen

March 1996

On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima

Nitish Shirish Keskar*
Northwestern University
Evanston, IL 60208
keskar.nitish@u.northwestern.edu

Dheevatsa Mudigere
Intel Corporation
Bangalore, India
dheevatsa.mudigere@intel.com

Jorge Nocedal
Northwestern University
Evanston, IL 60208
j-nocedal@northwestern.edu

Mikhail Smelyanskiy
Intel Corporation
Santa Clara, CA 95054
mikhail.smelyanskiy@intel.com

Ping Tak Peter Tang
Intel Corporation
Santa Clara, CA 95054
peter.tang@intel.com

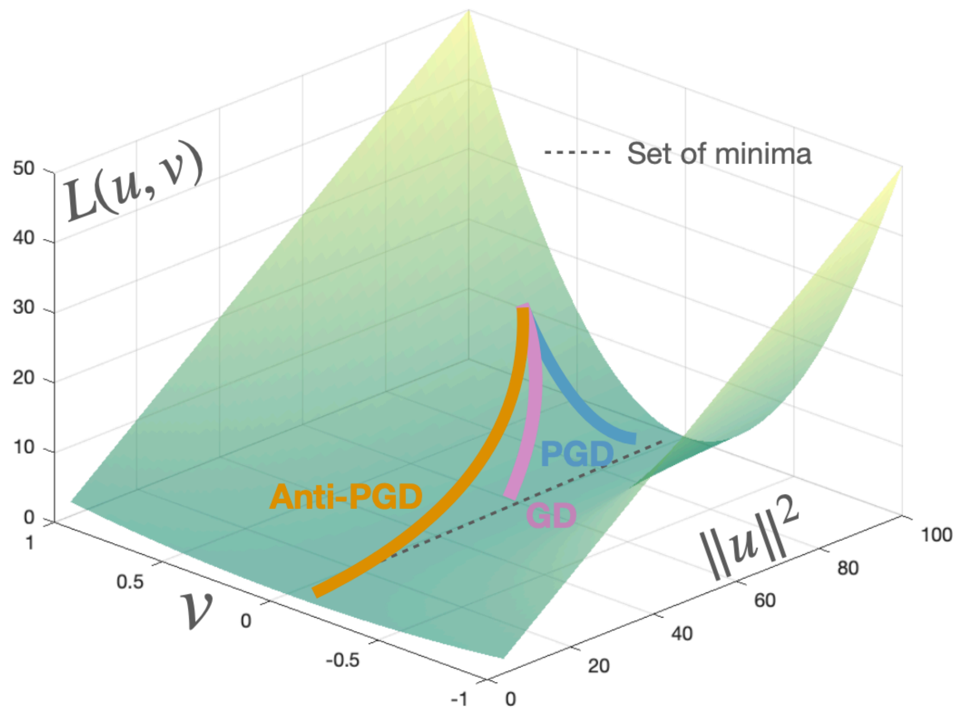
Exploring Generalization in Deep Learning

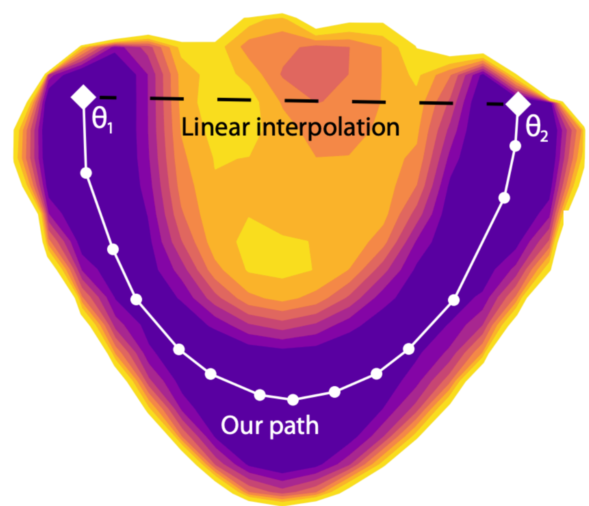
Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, Nathan Srebro
Toyota Technological Institute at Chicago
{bneyshabur, srinadh, mcallester, nati}@ttic.edu

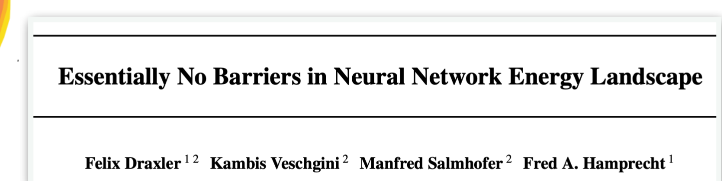
Why does it work? (2)

➡ We perform exact computations for a “widening valley” loss

$$L(u, v) = \frac{1}{2} v^2 \|u\|^2 \quad v \in \mathbb{R}, \text{ and } u \in \mathbb{R}^d$$



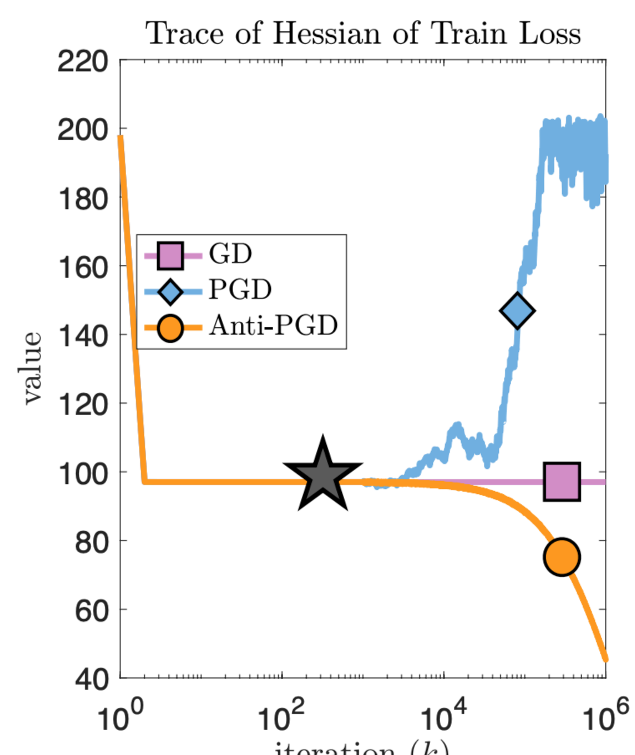


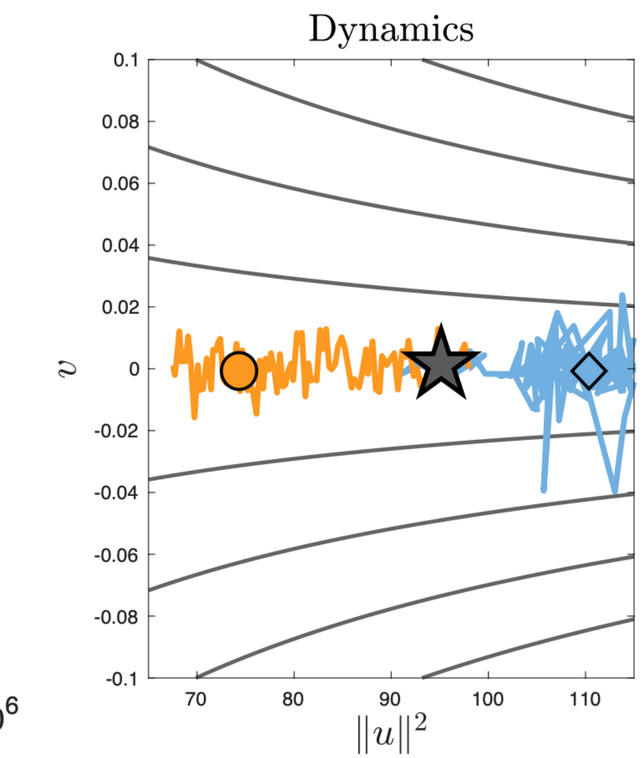

Essentially No Barriers in Neural Network Energy Landscape

Felix Draxler [1,2] Kambis Veschgini [2] Manfred Salmhofer [2] Fred A. Hamprecht [1]

See Theorem 3.1 in our paper!

- Anti-PGD converges to 0 (wide), while PGD diverges to sharp minima.
- Hyperparameter tuning does not help PGD.





Also check out our follow-up preprint!

Explicit Regularization in Overparametrized Models via Noise Injection

Antonio Orvieto*
Department of Computer Science
ETH Zürich, Zürich, CH.
antonio.orvieto@inf.ethz.ch

Anant Raj*
Coordinated Science Laboraotry
University of Illinois Urbana-Champaign.
Inria, Ecole Normale Supérieure
PSL Research University, Paris, France.
anant.raj@inria.fr

Hans Kersting*
Inria, Ecole Normale Supérieure
PSL Research University, Paris, France.
hans.kersting@inria.fr

Francis Bach
Inria, Ecole Normale Supérieure
PSL Research University, Paris, France.
francis.bach@inria.fr

June 13, 2022

Thank you!