Theoretical Advances in Deep Learning
JSM 2020 Discussion

Edgar Dobriban
Wharton, UPenn

August 4, 2020

Thanks!

- Po-Ling (organizing, chairing, invitation, ...)
- Jason, Tengyuan, Matus (talks)
- Dan, Misha (discussion)
- Audience (time and attention)
- Let's do this again next year!

Talks

- Good Linear Classifiers Are Abundant in the Interpolating Regime
  Jason Klusowski, Ryan Theisen, Michael Mahoney
- Multiple Descent Phenomenon, Risk of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels
  Tengyuan Liang, Pragya Sur
- Polylogarithmic Width Suffices for Gradient Descent to Achieve Arbitrarily Small Test Error with Shallow ReLU Networks
  Matus Telgarsky, Ziwei Ji

Good Linear Classifiers Are Abundant in the Interpolating Regime

- main messages: "worst-case" analysis too loose to "explain deep learning"; average case/typical analysis à la stat phys 1980s+ may be more insightful (DJ Amit, H Sompolinsky, E Gardner, M Mezard, B Derrida, N Tishby, G Gyorgyi, M Opper, HS Seung ...)
- "nearly all" interpolating classifiers are "good"
- Def: "test error distribution" of classifiers, fixed test/train data
- Leverage Linear Elliptical Slice Sampling [Gessner et al, 2020], to compute this on MNIST-scale
- Theory for equicorrelated data.

On the history of "average case" analysis high-dim classif

3.5. *Literature review for high-dimensional RDA.* There has been substantial work in the former Soviet Union on high-dimensional classification; references on this work include Raudys and Young (2004), Raudys (2001), and Serdobolskii (2007). Raudys (1967) derived the $n, p \to \infty$ asymptotic error rate of independence rules in identity-covariance case $\Sigma = I_p$, while Deev (1970) and Raudys (1972) obtained the error rate of unregularized linear discriminant analysis (LDA) for general covariance Σ, again in the $n, p \to \infty$ regime. A difference is that Raudys (1972) establishes normality of the linear discriminant function, whereas Deev (1970) expands the conditional probability of misclassification.

Serdobolskii (2007) calls the framework $n, p \to \infty$, $p/n \to \gamma$ the "Kolmogorov asymptotic regime," and suggests that around 1967 Kolmogorov was interested in this area. As explained by one of our referees, Kolmogorov had suggested the problem of studying Fisher's LDA under $n, p \to \infty$ asymptotics to Y. Blagovechenskij and his PhD student, A. Deev.

Figure: from Dobriban & Wager (2018, AoS).

On the history of "average case" analysis high-dim classif

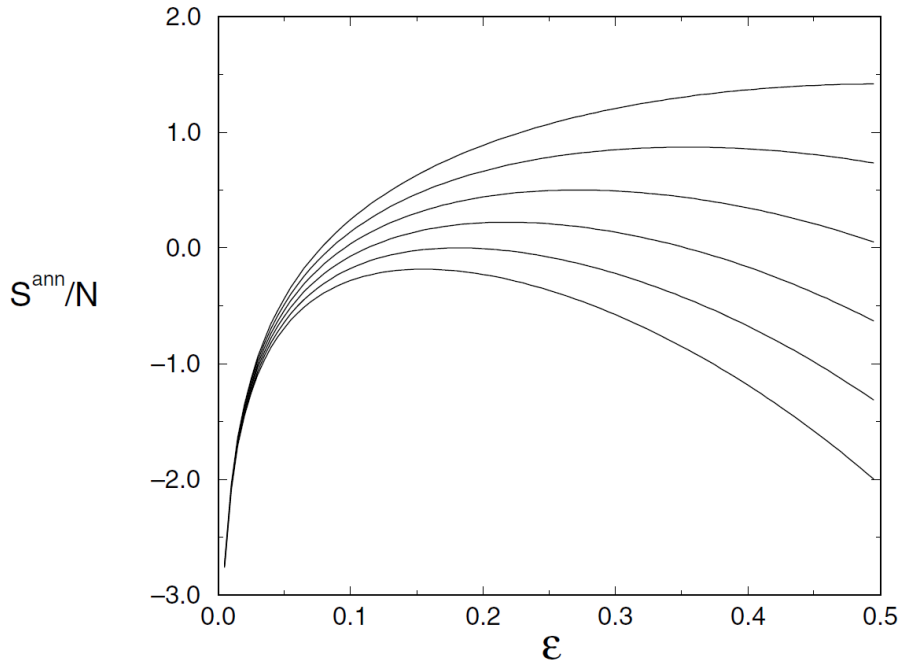

Fig. 2.2. Expression in the square brackets of (2.6) as a function of ε for $\alpha = 0, 1, 2, 3, 4$ and 5 (from top to bottom)

$$\Omega_p(\varepsilon) \sim \exp\left(N \left[\frac{1}{2}(1 + \ln 2\pi) + \frac{1}{2} \ln \sin^2(\pi\varepsilon) + \alpha \ln(1 - \varepsilon) \right] \right). \quad (2.6)$$

Figure: from Engel, Van den Broeck, Stat Mech of Learning, 2001

- G Gyorgyi, N Tishby. Statistical theory of learning a rule. Workshop on NNs and Spin Glasses, 1990.

Multiple Descent Phenomenon, Risk of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels

- Continuing recent line of work in double descent/interpolation (M Belkin, D Hsu, P Bartlett, P Long, A Montanari, S Mei, RJ Tibshirani, T Liang, A Rakhlin, A Sahai, ...)
- min-norm interpolation in RKHS, $d \propto n^{\alpha}$, $\alpha < 1$ can have "peaks" at $\alpha = 1/i$, highly intriguing...
- Universality?
- Regularization? Belkin, Hsu, Ma, Mandal: "Reconciling modern machine learning ..." (PNAS, 2019)

  "Regularization, of all forms, can both prevent interpolation and change the effective capacity of the function class, thus attenuating or masking the interpolation peak."

Optimal Regularization Can Mitigate Double Descent

Optimal Regularization Can Mitigate Double Descent

Preetum Nakkiran[1], Prayaag Venkat[1], Sham Kakade[2], and Tengyu Ma[3]

[1]Harvard University
[2]Microsoft Research & University of Washington
[3]Stanford University

Abstract

Recent empirical and theoretical studies have shown that many learning algorithms – from linear regression to neural networks – can have test performance that is non-monotonic in quantities such the sample size and model size. This striking phenomenon, often referred to as "double descent", has raised questions of if we need to re-think our current understanding of generalization. In this work, we study whether the double-descent phenomenon can be avoided by using optimal regularization. Theoretically, we prove that for certain linear regression models with isotropic data distribution, optimally-tuned ℓ_2 regularization achieves monotonic test performance as we grow either the sample size or the model size. We also demonstrate empirically that optimally-tuned ℓ_2 regularization can mitigate double descent for more general models, including neural networks. Our results suggest that it may also be informative to study the test risk scalings of various algorithms in the context of appropriately tuned regularization.

Asymptotic mitigation of DD, Dobriban & Sheng

Journal of Machine Learning Research 21 (2020) 1-52 Submitted 4/19; Revised 3/20; Published 4/20

WONDER: Weighted One-shot Distributed Ridge Regression in High Dimensions

Edgar Dobriban DOBRIBAN@WHARTON.UPENN.EDU
Wharton Statistics Department
University of Pennsylvania
Philadelphia, PA 19104, USA

Yue Sheng YUESHENG@SAS.UPENN.EDU
Graduate Group in Applied Mathematics and Computational Science
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor: John Shawe-Taylor

Proposition 6 (Properties of the optimal risk function) *The optimal risk function* $\phi(\gamma)$ *has the following properties:*

1. **Monotonicity:** $\phi(\gamma)$ *is an increasing function of* $\gamma \in [0, \infty)$ *with* $\lim_{\gamma \to 0+} \phi(\gamma) = 0$ *and* $\lim_{\gamma \to +\infty} \phi(\gamma) = \alpha^2$.

Figure: The asymptotic risk of optimally regularized ridge in white noise is monotone increasing as a function of $\gamma = \lim p/n$. This suggests that the risk is (1) decreasing for "fixed" n and decreasing p (consistent with double descent being mitigated), and (2) decreasing for "fixed" p and increasing n (consistent with sample-wise double descent being mitigated).

History of DD

- Loog et al: A Brief Prehistory of Double Descent, PNAS 2020 letter
- Belkin et al: Looking beyond the peaking phenomenon, PNAS 2020 letter
- Twitter thread by Dmitry Kobak `twitter.com/hippopedoid/status/1243229021921579010`. Points to works by Hertz et al. [1989], Opper et al. [1990], Hansen [1993], Barber et al. [1995], Duin [1995], Opper [1995], Opper and Kinzel [1996], Raudys and Duin [1998]

Polylogarithmic Width Suffices for Gradient Descent to Achieve Arbitrarily Small Test Error with Shallow ReLU Networks

- One of the latest in long line of work on provable non-convex optimization+generalization for $\geqslant 2$ layer NNs
- Deeply leverages
  - Homogeneity: $\langle W, \nabla_W f(W) \rangle = f(W)$
  - "Self-bounding" property of logistic loss $\ell(z) = \log(1 + \exp(-z))$: $-\ell' \leqslant \ell$
  - Generalization error can be bounded by bounded+Lipschitz surrogate: $I(z \leqslant 0) \leqslant -2\ell'(z)$

Application to data augmentation

A Group-Theoretic Framework for Data Augmentation

Shuxiao Chen, Edgar Dobriban* and Jane H Lee†

February 25, 2020

Abstract

Data augmentation is a widely used trick when training deep neural networks: in addition to the original data, properly transformed data are also added to the training set. However, to the best of our knowledge, a clear mathematical framework to explain the performance benefits of data augmentation is not available.

In this paper, we develop such a theoretical framework. We show data augmentation is equivalent to an averaging operation over the orbits of a certain group that keeps the data distribution approximately invariant. We prove that it leads to variance reduction. We study empirical risk minimization, and the examples of exponential families, linear regression, and certain two-layer neural networks. We also discuss how data augmentation could be used in problems with symmetry where other approaches are prevalent, such as in cryo-electron microscopy (cryo-EM).

Application to data augmentation

Theorem 7.1 (Benefits of data augmentation for two-layer ReLU nets)**.** *Under Assumption C, take any* $\varepsilon \in (0,1)$ *and* $\delta \in (0, 1/5)$*. Let*

$$\lambda = \frac{\sqrt{2\log(4n|G|/\delta)} + \log(4/\varepsilon)}{\gamma/4}, M = \frac{4096\lambda^2}{\gamma^6},$$

and let $\rho = 4\lambda/(\gamma\sqrt{m})$*. Let* k *be the best iteration (with the lowest empirical risk) in the first* $\lceil 2\lambda^2/n\varepsilon \rceil$ *steps. Let* $\alpha = 16[\sqrt{2\log(4n|G|/\delta)} + \log(4/\varepsilon)]/\gamma^2 + \sqrt{md} + \sqrt{2\log(1/\delta)}$*. For any* $m \geq M$ *and any constant step size* $\eta \leq 1$*, with probability at least* $1 - 5\delta$ *over the random initialization and i.i.d. draws of the data points, we have*

$$\mathbb{P}(Y f(X; W_k, a) \leq 0) \leq 2\varepsilon + [\sqrt{\frac{2\log 2/\delta}{n}} + 4\bar{\mathcal{R}}_n] + \frac{1}{2}\mathbb{E}_Y \mathbb{E}_g \mathcal{W}_1(X|Y, gX|Y) \cdot \alpha.$$

The three terms bound the optimization error, generalization error, and the bias due to approximate invariance. Moreover, with probability at least $1 - \delta$ *over the random initialization, we have*

$$\bar{\mathcal{R}}_n - \mathcal{R}_n \leq \Delta + \frac{1}{4}\mathbb{E}_Y \mathbb{E}_g \mathcal{W}_1(X|Y, gX|Y) \cdot \alpha,$$

where

$$\Delta = \mathbb{E} \sup_{W \in \mathscr{W}_\rho} |\frac{1}{n}\varepsilon_i \mathbb{E}_g[-\ell'(y_i f_{i,g}(W))]| - \mathbb{E}\mathbb{E}_g \sup_{W \in \mathscr{W}_\rho} |\frac{1}{n}\varepsilon_i[-\ell'(y_i f_{i,g}(W))]| \leq 0$$

is the "variance reduction" term.

Proof. See Appendix B.5. □

The proof idea is largely based on results in Ji and Telgarsky (2019). We decompose the overall error into optimization error and generalization error. The optimization error is taken care of by a corollary of Theorem 2.2 in Ji and Telgarsky (2019). The generalization error is dealt with by adapting several arguments in Theorem 3.2 of Ji and Telgarsky (2019) and using some arguments in the proof of Theorem 6.4.

References I

David Barber, David Saad, and Peter Sollich. Finite-size effects and optimal test set size in linear perceptrons. *Journal of Physics A: Mathematical and General*, 28(5):1325, 1995.

Robert PW Duin. Small sample size generalization. In *Proceedings of the Scandinavian Conference on Image Analysis*, volume 2, pages 957–964. PROCEEDINGS PUBLISHED BY VARIOUS PUBLISHERS, 1995.

Lars Kai Hansen. Stochastic linear learning: Exact test and training error averages. *Neural Networks*, 6(3):393–396, 1993.

JA Hertz, A Krogh, and GI Thorbergsson. Phase transitions in simple learning. *Journal of Physics A: Mathematical and General*, 22(12):2133, 1989.

M Opper, W Kinzel, J Kleinz, and R Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11):L581, 1990.

Manfred Opper. Statistical mechanics of learning: Generalization. *The Handbook of Brain Theory and Neural Networks,*, pages 922–925, 1995.

References II

Manfred Opper and Wolfgang Kinzel. Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer, 1996.

Sarunas Raudys and Robert PW Duin. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern recognition letters*, 19(5-6):385–392, 1998.