Generalization Error of Generalized Linear Models in High Dimensions

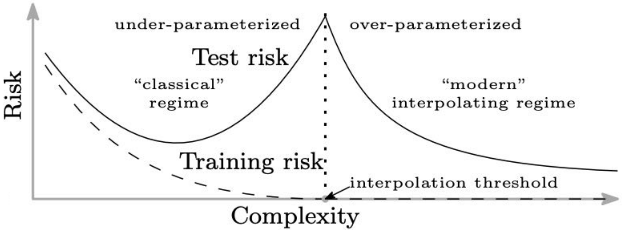
Melika Emami [1] , Mojtaba Sahraee-Ardakan [1,2],
Parthe Pandit[1,2], Sundeep Rangan[3], Alyson K. Fletcher [1,2]

[1]ECE, UCLA, [2]STAT, UCLA, [3]ECE, NYU

ICML 2020

Overview

- Generalization Error: Performance on new data
- Fundamental question in modern systems:
  - Low generalization error despite over-parameterization

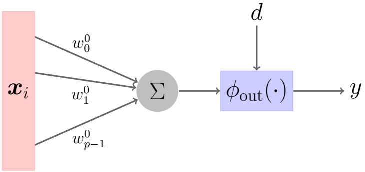


[BHMM19]

- **This work**: Exact calculation of generalization error for GLMs
  - High dimensional regime
  - *Double descent* phenomenon

Overview

- Generalized linear models (GLMs): $y = \phi_{\text{out}}(\langle \boldsymbol{x}, \boldsymbol{w}^0 \rangle, d)$



- Regularized ERM:

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\text{argmin}} \quad F_{\text{out}}(\boldsymbol{y}, \mathbf{X}\boldsymbol{w}) + F_{\text{in}}(\boldsymbol{w})$$

- Generalization error:

$$\mathbb{E}\, f_{\text{ts}}(y_{\text{ts}}, \widehat{y}_{\text{ts}}) \tag{1}$$

  - Test sample: $(\boldsymbol{x}_{\text{ts}}, y_{\text{ts}})$
  - $y_{\text{ts}} = \phi_{\text{out}}(\left\langle \boldsymbol{x}_{\text{ts}}, \boldsymbol{w}^0 \right\rangle, d_{\text{ts}}), \quad \widehat{y}_{\text{ts}} = \phi(\langle \boldsymbol{x}_{\text{ts}}, \widehat{\boldsymbol{w}} \rangle)$

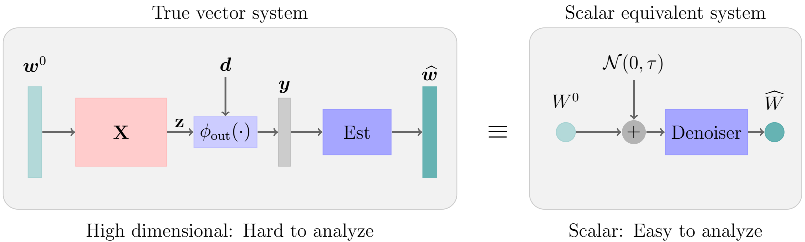
Overview

- Prior work
  - Understanding generalization in deep neural nets [BMM18, BHX19, BLLT19, NLB+18, ZBH+16, AS17]
  - Linear models [MRSY19, DKT19, MM19, HMRT19, GAK20]
  - GLMs with uncorrelated features [BKM+19]
- **Our contribution**:
  - A procedure for characterizing generalization error (1)
  - General test metrics, training losses, regularizers, link function
  - Correlated covariates
  - Train-test distributional mismatch
  - Over-parameterized and under-parameterized regime

Outline

- Main Result
  - Scalar Equivalent System
  - Main Theorem
- Examples
  - Linear Regression
  - Logistic Regression
  - Non-linear Regression
- Proof Technique
  - Multi-layer VAMP
- Future Directions

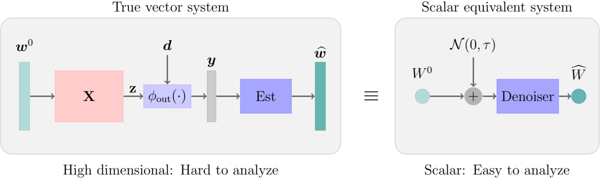
Scalar Equivalent System



$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\text{argmin}} \quad F_{\text{out}}(\boldsymbol{y}, \mathbf{X}\boldsymbol{w}) + F_{\text{in}}(\boldsymbol{w}) \tag{2}$$

- **Key tool**: Approximate Message Passing (AMP) framework [DMM09, BM11, RSF19, FRS18, PSAR+20]
  - As a constructive proof technique
  - Performance of the estimates:
    $\rightarrow$ deterministic recursive equations: *state evolution* (SE)

Main Result



Theorem (Generalization error of GLMs)

(a) Under some regularity conditions on $f_{\text{ts}}, \phi, \phi_{\text{out}}$, the above convergence is rigorous:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(w_i^0, \widehat{w}_i) = \mathbb{E} f(W^0, \widehat{W}) \quad a.s.$$

$$\widehat{W} = \text{prox}_{f_{\text{in}}/\gamma}(W^0 + Q), \quad Q = \mathcal{N}(0,\tau) \quad (\textit{independent of } W^0)$$

(b) Generalization error:

$$\mathcal{E}_{\text{ts}} = \mathbb{E}\, f_{\text{ts}}\left(\phi_{\text{out}}(Z_{\text{ts}}, D), \phi(\widehat{Z}_{\text{ts}})\right), \quad (Z_{\text{ts}}, \widehat{Z}_{\text{ts}}) \sim \mathcal{N}(\mathbf{0}_2, \mathbf{M})$$

τ, γ, *and* $\mathbf{M}$ *are computed by SE equations, and* $D \perp\!\!\!\perp (Z_{\text{ts}}, \widehat{Z}_{\text{ts}})$

Example Setting

- Train-test distributional mismatch
  - $\boldsymbol{x}_{\text{train}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{tr}})$, $\boldsymbol{x}_{\text{test}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{ts}})$, $\boldsymbol{\Sigma}_{\text{tr}}$ and $\boldsymbol{\Sigma}_{\text{ts}}$ commute
  - i.i.d. log-normal eigenvalues

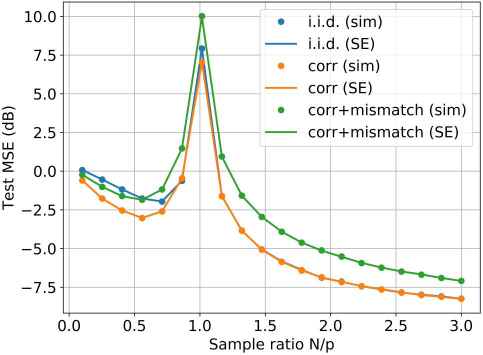
$$\begin{bmatrix} \log(S_{\text{tr}}^2) \\ \log(S_{\text{ts}}^2) \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \sigma \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \quad \forall\ i$$

- 3 different cases :

(i) Uncorrelated features ($\sigma = 0$) $\quad \boldsymbol{\Sigma}_{\text{tr}} = \boldsymbol{\Sigma}_{\text{ts}} = \boldsymbol{I}$
(ii) Correlated features ($\sigma > 0,\ \rho = 1$) $\quad \boldsymbol{\Sigma}_{\text{tr}} = \boldsymbol{\Sigma}_{\text{ts}} \neq \boldsymbol{I}$
(iii) Mismatched features ($\sigma > 0,\ \rho < 1$) $\quad \boldsymbol{\Sigma}_{\text{tr}} \neq \boldsymbol{\Sigma}_{\text{ts}}$

Example: Linear Regression

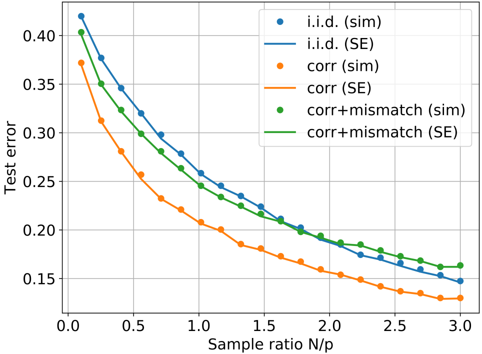
- Under-regularized linear regression:
  - $\phi_{\text{out}}(p, d) = p + d$, and $d \sim \mathcal{N}(0, \sigma_d^2)$
  - MSE output loss
  - *double descent* phenomenon



(Recovered result of [HMRT19])

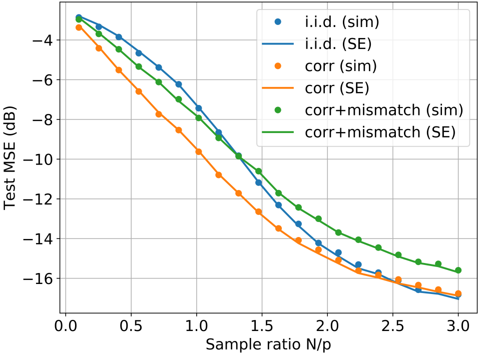
Example: Logistic Regression

- Logistic regression
  - Logistic output $P(y=1) = 1/(1+e^{-p})$
  - Binary cross-entropy loss with ℓ_2 regularization

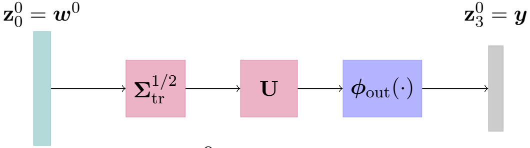


Example: Non-linear Regression

- Non-linear Regression
  - $\phi_{\text{out}}(p, d) = \tanh(p) + d, \quad d \sim \mathcal{N}(0, \sigma_d^2)$
  - $f_{\text{out}}(y, p) = \frac{1}{2\sigma_d^2}(y - \tanh(p))^2$



Proof Technique: Multi-Layer Representation



- Represent the mapping $\boldsymbol{w}^0 \mapsto \boldsymbol{y}$ as a multi-layer network

$$\boldsymbol{y} = \phi_{\mathrm{out}}(\mathbf{X}\boldsymbol{w}, \mathbf{d})$$

- Decompose Gaussian training data $\mathbf{X}$ with covariance $\boldsymbol{\Sigma}_{\mathrm{tr}}$

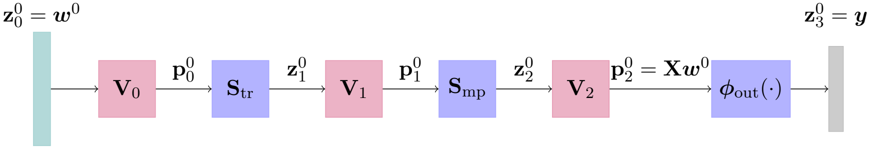
$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}_{\mathrm{tr}}^{\frac{1}{2}}, \quad \mathbf{U} \text{ i.i.d. Gaussian}$$

- Use SVD of $\mathbf{U}$ and eigendecomposition of $\boldsymbol{\Sigma}_{\mathrm{tr}}^{\frac{1}{2}}$:

$$\boldsymbol{\Sigma}_{\mathrm{tr}} = \frac{1}{p}\mathbf{V}_0^{\mathsf{T}}\mathrm{diag}(\mathbf{s}_{\mathrm{tr}}^2)\mathbf{V}_0, \quad \mathbf{U} = \mathbf{V}_2\mathbf{S}_{\mathrm{mp}}\mathbf{V}_1$$

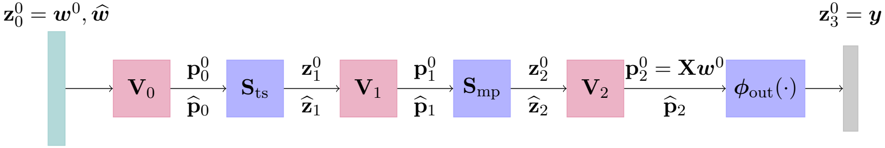
- $\mathbf{V}_0$, $\mathbf{V}_1$, $\mathbf{V}_2$: Haar-distributed
- $\mathbf{S}_{\mathrm{mp}}$: Singular values of $\mathbf{U}$
  - converges in distribution to Marchenko-Pastur law

Proof Technique: Multi-Layer VAMP



- Algorithm to solve inference problem in deep neural networks
- Similar to ADMM algorithm for optimization
- Statistical guarantees:
  - Joint distribution of $(W^0, \widehat{W})$ and other hidden signals

Proof Technique: Generalization Error



- ML-VAMP $\Rightarrow$ Joint distribution of $(W^0, \widehat{W})$ (part (a) of Thm)
- Given test data:

$$\boldsymbol{x}_{\text{ts}}^{\mathsf{T}} = \mathbf{u}^{\mathsf{T}}\text{diag}(\mathbf{s}_{\text{ts}})\mathbf{V}_0$$

- Find joint distribution of $(P_2^0, \widehat{P}_2)$ for test data (part (b) of Thm)

$$(P_2^0, \widehat{P}_2) \sim \mathcal{N}(\mathbf{0}_2, \mathbf{M})$$

- Obtain generalization error

$$\mathcal{E}_{\text{ts}} = \mathbb{E}\, f_{\text{ts}}\Big(\phi_{\text{out}}(P_2^0, D), \phi(\widehat{P}_2)\Big)$$

Future Directions

- Generalize results to:
  - Non-Gaussian covariates
  - Multitask GLMs using multi-layer matrix-valued VAMP
  - Deeper models like two-layer neural networks
  - Non-asymptotic regimes
- Use results to get:
  - Generalization errors in reproducing kernel Hilbert spaces, such as NTK space

Madhu S Advani and Andrew M Saxe.
High-dimensional dynamics of generalization error in neural networks.
arXiv preprint arXiv:1710.03667, 2017.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal.
Reconciling modern machine-learning practice and the classical bias–variance trade-off.
Proc. National Academy of Sciences, 116(32):15849–15854, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu.
Two models of double descent for weak features.
arXiv preprint arXiv:1903.07571, 2019.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová.
Optimal errors and phase transitions in high-dimensional generalized linear models.
Proc. National Academy of Sciences, 116(12):5451–5460, March 2019.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler.
Benign overfitting in linear regression.
arXiv preprint arXiv:1906.11300, 2019.

M. Bayati and A. Montanari.
The dynamics of message passing on dense graphs, with applications to compressed sensing.
IEEE Trans. Inform. Theory, 57(2):764–785, February 2011.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal.
To understand deep learning we need to understand kernel learning.
arXiv preprint arXiv:1802.01396, 2018.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis.
A model of double descent for high-dimensional binary linear classification.
arXiv preprint arXiv:1911.05822, 2019.

David L Donoho, Arian Maleki, and Andrea Montanari.

Message-passing algorithms for compressed sensing.
Proc. National Academy of Sciences, 106(45):18914–18919, 2009.

Alyson K Fletcher, Sundeep Rangan, and P. Schniter.
Inference in deep networks in high dimensions.
Proc. IEEE Int. Symp. Information Theory, 2018.

Cédric Gerbelot, Alia Abbara, and Florent Krzakala.
Asymptotic errors for convex penalized linear regression beyond gaussian matrices.
arXiv preprint arXiv:2002.04372, 2020.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani.
Surprises in high-dimensional ridgeless least squares interpolation.
arXiv preprint arXiv:1903.08560, 2019.

Song Mei and Andrea Montanari.
The generalization error of random features regression: Precise asymptotics and double descent curve.
arXiv preprint arXiv:1908.05355, 2019.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan.
The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime.
arXiv preprint arXiv:1911.01544, 2019.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro.
Towards understanding the role of over-parametrization in generalization of neural networks.
arXiv preprint arXiv:1805.12076, 2018.

Parthe Pandit, Mojtaba Sahraee-Ardakan, Sundeep Rangan, Philip Schniter, and Alyson K Fletcher.
Inference with deep generative priors in high dimensions.
IEEE Journal on Selected Areas in Information Theory, 2020.

Sundeep Rangan, Philip Schniter, and Alyson K Fletcher.
Vector approximate message passing.
IEEE Trans. Information Theory, 65(10):6664–6684, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals.
Understanding deep learning requires rethinking generalization.
arXiv preprint arXiv:1611.03530, 2016.