High-Dimensional Analysis for Generalized Nonlinear Regression: From Asymptotics to Algorithm

Jian Li[1], Yong Liu[2,*] and Weiping Wang[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences

[2] Gaoling School of Artificial Intelligence, Renmin University of China

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Backgrounds

Related works:

- Benign overfitting: Overparameterized models often achieve benign overfitting, interpolating the training data while still generalizing well.
- Double descent phenomenon: The testing error characterizes a U-shaped performance curve in the under-parameterized regime, while it decreases again in the over-parameterized regime.

Motivation: Despite the extensive literature devoted to understanding the double descent phenomenon, there are still several open problems:

1) The lack of a general asymptotic analysis framework for generalized nonlinear regression models.
2) Existing asymptotic results often remain as self-consistency equations that are hard to estimate.
3) Benign overfitting can be caused by overparameterization, and subsampling may also achieve better performance from a dual view.

Contributions:

- Generalized asymptotic analysis framework for nonlinear regression models.
- Trainable nonlinear regression algorithm based on theoretical findings.
- Interesting byproducts: the use of nonlinear feature mapping to reduce effective dimension and the potential benefits of subsampling for generalization.

Preliminaries

- Linear Ridge Regression

$$\operatorname*{arg\,min}_{\eta\in\mathbb{R}^d}\left\{\frac{1}{n}\sum_{i=1}^n\left(\eta^\top x-y_i\right)^2+\lambda\|\eta\|_2^2\right\},\quad \text{with}$$

$$\hat\eta=(\widehat\Sigma+\lambda I)^{-1}\widehat\Sigma\eta_*+(\widehat\Sigma+\lambda I)^{-1}\frac{X^\top\varepsilon}{n}.$$

where $\widehat\Sigma=\frac{1}{n}X^\top X\in\mathbb{R}^{d\times d}$ the covariance matrix.

- Generalized Nonlinear Regression Model

$$\operatorname*{argmin}_{\theta\in\mathbb{R}^p}\left\{\frac{1}{n}\|\phi(X)\theta-y\|_2^2+\lambda\|\theta\|_2^2\right\},\quad \text{with}$$

$$\hat\theta=(\widehat\Sigma_\phi+\lambda I)^{-1}\widehat\Sigma_\phi\theta_*+(\widehat\Sigma_\phi+\lambda nI)^{-1}\frac{\phi(X)^\top\varepsilon}{n},$$

where $\phi:\mathbb{R}^d\to\mathbb{R}^p$ is the feature mapping.

- Nonlinear Regression Model with Subsampling

$$\operatorname*{argmin}_{\theta\in\mathbb{R}^p}\left\{\frac{1}{m}\|S\phi(X)\theta-Sy\|_2^2+\lambda\|\theta\|_2^2\right\},\quad \text{with}$$

$$\hat\theta=\left(\widehat\Sigma_{S\phi}+\lambda I\right)^{-1}\widehat\Sigma_{S\phi}\theta_*+\left(\widehat\Sigma_{S\phi}+\lambda I\right)^{-1}\frac{\phi(X)^\top S^\top S\varepsilon}{m},$$

Assumptions

Assumption 1 (Existence of θ_* in the feature space).

Assumption 2 (Continuous and bounded feature mapping).

Assumption 3 (Covariance condition for nonlinear feature mapping). *Supoose Σ_ϕ is invertible and bounded, and the eigenvalues of Σ_ϕ are positive and bounded. $\phi(X)=Z\Sigma_\phi^{1/2}$ where Z has i.i.d. entries with zero mean, and unit variance.*

Assumption 4 (Orthogonal subsampling matrix). *Suppose the rows of subsampling matrix is orthogonal, such that $SS^\top=I_m$. Meanwhile, $S^\top S$ converges to a deterministic matrix Σ_S.*

Assumption 5 (Covariance condition for subsampled nonlinear models). *The empirical covariance matrix of $\widehat\Sigma_{S\phi}=\frac{1}{m}\phi(X)^\top S^\top S\phi(X)$ converges to a deterministic covariance matrix $\Sigma_{S\phi}=\Sigma_\phi^{1/2}Z^\top\Sigma_S Z\Sigma_\phi^{1/2}$. The spectral distribution $F_{\Sigma_{S\phi}}$ of $\Sigma_{S\phi}$ converges to a limit probability distribution μ supported on $[0,+\infty)$ and Σ is invertible and bounded in operator norm.*

Asymptotics Results

Theorem 1 (Asymptotic risk for ridge regression). *Under Assumptions 2 - 5, the nonlinear ridge regression with subsampling estimator in (??) admits the following limiting variance and bias:*

$$\mathbb{E}_\varepsilon\left[\left\|\hat\theta-\mathbb{E}_\varepsilon(\hat\theta)\right\|_{\Sigma_{S\phi}}^2\right]\sim\sigma^2\frac{\mathrm{df}_2(\kappa)}{m-\mathrm{df}_2(\kappa)},$$

$$\left\|\mathbb{E}_\varepsilon(\hat\theta)-\theta_*\right\|_{\Sigma_{S\phi}}^2\sim\frac{m\kappa^2\theta_*^\top(\Sigma_{S\phi}+\kappa I)^{-2}\Sigma_{S\phi}\theta_*}{m-\mathrm{df}_2(\kappa)}.$$

Corollary 1 (Under-parameterized regime). *Under Assumptions 2 - 5, if $\lambda=0$ and $\gamma<1$, the nonlinear ridgeless regression with subsampling estimator admits:*

$$\mathbb{E}_\varepsilon\left[\left\|\hat\theta-\mathbb{E}_\varepsilon(\hat\theta)\right\|_{\Sigma_{S\phi}}^2\right]\sim\sigma^2\tfrac{p}{m-p},\qquad\left\|\mathbb{E}_\varepsilon(\hat\theta)-\theta_*\right\|_{\Sigma_{S\phi}}^2=0.$$

Corollary 2 (Over-parameterized regime). *Under Assumptions 2 - 5, if $\lambda=0$ and $\gamma>1$, with κ_0 defined by $\mathrm{df}_1(\kappa_0)=$ m the nonlinear ridgeless regression with subsampling estimator admits:*

$$\mathbb{E}_\varepsilon\left[\left\|\hat\theta-\mathbb{E}_\varepsilon(\hat\theta)\right\|_{\Sigma_{S\phi}}^2\right]\sim\sigma^2\frac{\mathrm{df}_2(\kappa_0)}{m-\mathrm{df}_2(\kappa_0)},$$

$$\left\|\mathbb{E}_\varepsilon(\hat\theta)-\theta_*\right\|_{\Sigma_{S\phi}}^2=\frac{m\kappa^2\theta_*^\top(\Sigma_{S\phi}+\kappa I)^{-2}\Sigma_{S\phi}\theta_*}{m-\mathrm{df}_2(\kappa_0)}.$$

Algorithm: RFRed

Based on random Fourier features, we devise Random Feature Regression model with Effective Dimension (RFRed)

$$\phi(x)=\sqrt{\frac{2}{p}}\cos(W^\top x+b),$$

where the frequency matrix $W=[w_1,\cdots,w_p]\in\mathbb{R}^{d\times p}$ is trainable and initialized by a Gaussian distribution. The phase vectors $b=[b_1,\cdots,b_p]\in\mathbb{R}^p$ are drawn uniformly from $[0,2\pi]$.
Motivated by the asymptotics results, we devise the following objective and optimize θ and W jointly.

$$\mathcal{L}(\theta;\,W)=\frac{1}{n}\|S\phi(X)\theta-Sy\|_2^2+\lambda\|\theta\|_2^2+\beta\,\widehat{\mathrm{df}}_2(\lambda),$$

Complexity. Using batch stochastic gradient method, we have $\nabla_\theta\mathcal{L}=\frac{1}{n}\widetilde X_b^\top(\widetilde X_b\theta-\widetilde y_b)$ where $\{\widetilde X_b\in\mathbb{R}^{b\times p},\widetilde y_b\in\mathbb{R}^b\}$ is a batch of $\{S\phi(X),Sy\}$ with the batch size b. We also use the batch data to approximate $\widehat{\mathrm{df}}_2(\lambda)$ where $\widetilde X$ in (??) is replaced by $\widetilde X_b$. The compute of $S\phi(X)$ consumes $\mathcal{O}(mnp+ndp)$. With T iterations, the update of θ takes $\mathcal{O}(pbT)$ time, the update of W consumes $\mathcal{O}(pb^2T)$, and the compute of $\widehat{\mathrm{df}}_2(\lambda)$ requires $\mathcal{O}(\frac{p^2nT}{n\alpha})$.

Experiments