A Generalization Gap Estimation for Overparameterized Models via the Langevin Functional Variance (JCGS2023)

Akifumi Okuno[1,2,3] Keisuke Yano[1,2]

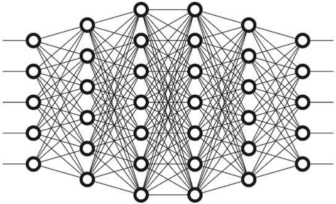
[1]ISM [2]SOKENDAI, [3]RIKEN

Abstract

The functional variance used as the penalty term in WAIC can estimate the generalization error of singular models with fixed parameter size, but its applicability to over-parameterized models such as neural networks remains unclear. In this study, we consider over-parameterized linear regression as a linearized setting and show that the functional variance becomes an asymptotically unbiased estimator of the generalization error. We also propose a method for estimating the functional variance using Langevin dynamics.

Overparameterized model g_θ (e.g., deep neural network):

$$y_i \approx g_\theta(\mathbf{z}_i), \quad \mathbf{z}_i \in \mathbb{R}^q, \quad (i = 1, 2, \ldots, n).$$



Its linear approximation (called *overparameterized linear regression*; Bartlett et al., 2020)

$$y_i \approx \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle, \quad \mathbf{x}_i, \boldsymbol{\beta} \in \mathbb{R}^p, \quad (i = 1, 2, \ldots, n)$$

satisfies $p \approx$ "$\sharp$parameters in g_θ". Namely, $n \leq p$.

(Gibbs) Generalization Gap

$$\Delta(\alpha) = \underbrace{\mathbb{E}_{\mathbf{y}^*,\mathbf{y}}\left(\mathbb{E}_{\boldsymbol{\beta}\sim\mathrm{Pos}(\alpha)}[\|\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}\|_2^2]\right)}_{\text{for test}} - \underbrace{\mathbb{E}_{\mathbf{y}}\left(\mathbb{E}_{\boldsymbol{\beta}\sim\mathrm{Pos}(\alpha)}[\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2]\right)}_{\text{for training}}$$

with a ridge-estimator $\hat{\boldsymbol{\beta}}_\alpha = (\mathbf{X}^\top \mathbf{X}/n + \alpha)^{-1}\mathbf{X}^\top \mathbf{y}/n$ and a quasi-posterior

$$\mathrm{Pos}(\alpha) = N(\hat{\boldsymbol{\beta}}_\alpha, \mathbf{Q}_\alpha).$$

- **Cross-validation**:
  - 👎 requires retraining (=computationally intensive)
- **Information Criterion**:
  - 👎 cannot be applied to singular models (e.g., DNN)
  - 👎 requires $p \times p$ information matrix (=computationally intensive)

Functional Variance

$$\mathrm{FV}(\alpha) = \sum_{i=1}^{n} \mathbb{V}_{\boldsymbol{\beta} \sim \mathrm{Pos}(\alpha)}[\log f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta})]$$

is a penalty term in WAIC (Watanabe, 2010, 2018).

- 👍 $|\mathbb{E}_{\mathbf{y}}[\mathrm{FV}(\alpha)] - \Delta(\alpha)| \to^{p} 0$ even for singular models (with p:fixed, $n \to \infty$).

Two problems:

- 👎 *(Theory)* p in overparameterized models is not fixed ($p \geq n$).
- 👎 *(Computation)* Posterior computation for NN is difficult.

★ Our Contribution 1 (Theory)

Main Theorem (Informal)

Let $p \geq n$. Under the following conditions

(i) $\max_{\mathbf{u},i} \mathbb{P}(\mathbf{u} :=$ i-th left singular vector of $\mathbf{X}_n) = o(n/\mathrm{vol}(\mathbb{S}^{n-1}))$,

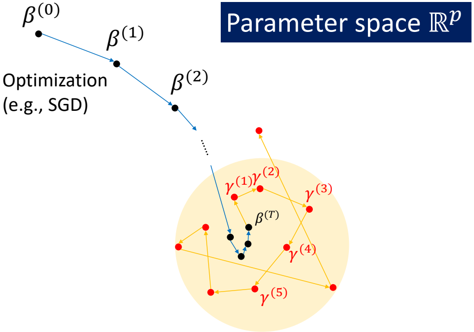
(ii) $\mathbb{P}\Bigg(\underbrace{\frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbf{X}_n)^2}_{\approx n^{-1}\sum_{i=1}^{n} \lambda_i(\mathrm{Fisher}(g_\theta))} < s_* \Bigg) \to 1 \; (n \to \infty)$, and

(iii) $\|\boldsymbol{\beta}_0\|_\infty \leq p^{-1/2} b$,

it holds that $\quad |\mathbb{E}_{\mathbf{y}}[\mathrm{FV}(\alpha)] - \Delta(\alpha)| \to^p 0 \quad (n \to \infty)$.

- 👍 Gaussian $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ $(\mathrm{tr}\boldsymbol{\Sigma}_n \leq \lambda_*)$ (see, e.g., Eaton (1989))
- 👍 Deep/Shallow NN g_θ (see, e.g., Karakida et al. (2019))

Langevin Gradient Descent



Posterior ≈ Langevin Distribution

$$\boldsymbol{\gamma}^{(t+1)} = \underbrace{\boldsymbol{\gamma}^{(t)} - \frac{1}{4}\delta \frac{n}{\sigma_0^2} \frac{\ell_\alpha(\boldsymbol{\gamma}^{(t)})}{\partial \boldsymbol{\gamma}}}_{\text{Gradient Descent}} + \underbrace{\delta^{1/2}\mathbf{e}^{(t)}}_{\text{Normal Noise}}, \quad \mathbf{e}^{(t)} \sim N_p(\mathbf{0}, \mathbf{I}_p).$$

★ Our Contribution 2 (Computation)

cf. Existing FV : $\quad \mathrm{FV}(\alpha) = \sum_{i=1}^{n} \hat{\mathbb{V}}_{\boldsymbol{\beta} \sim \mathrm{Pos}(\alpha)}[\log f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta})].$

Langevin FV : $\quad \mathrm{LFV}(\alpha) = \sum_{i=1}^{n} \hat{\mathbb{V}}_{\boldsymbol{\gamma} \sim \mathrm{Langevin}(\alpha)}[\log f(y_i \mid \mathbf{x}_i, \boldsymbol{\gamma})].$

- Simply computed with GD-based packages/wrappers (e.g., PyTorch)
- Does not compute $p \times p$ Information matrix (unlike TIC/RIC/GIC...)

Numerical Experiments 1: Linear

Synthetic Data Generation:

- $p = 2n$, with $n = 100, 200, 300, 400$.
- $\mathbf{U} \in \mathbb{R}^{n \times n}, \mathbf{V} \in \mathbb{R}^{n \times p}$ are uniformly sampled from orthogonal matrices
- $\boldsymbol{\beta}_0 \sim N(\mathbf{0}, p^{-1}\mathbf{I}_p)$,
- with given singular values $\mathbf{s} = \mathrm{diag}(s_1, s_2, \ldots, s_n)$,

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top \quad \text{and} \quad \mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}_0, \mathbf{I}_n).$$

LFV Setting:

- $\delta = 1/(10n)$, $T = 15n$.

Numerical Experiments 1-1: Linear ($s_i = \mathbb{1}(i \leq 10)$)

Table: $p/n = 2$, $\alpha = 0.1$ $\delta = 0.1/n$, $\sharp$Posterior $=$ $\sharp$Langevin $= 15n$, $\sharp$MonteCarlo $= 50$

n	100	200	300
$\Delta(\alpha)$	9.091	9.091	**9.091**
TIC($\kappa = 0$)	8.770 ± 1.412	9.473 ± 1.035	9.466 ± 0.885
TIC($\kappa = 0.1$)	8.770 ± 1.412	9.473 ± 1.035	9.466 ± 0.885
FV(α)	8.458 ± 1.286	8.845 ± 0.941	8.789 ± 0.802
LFV(α)	8.477 ± 1.368	8.966 ± 0.994	**8.917** ± 0.792

*TIC with generalized inverse of information matrix (Thomas et al., 2020)

Numerical Experiments 1-2: Linear ($s_i = i^{-1}$)

Table: $p/n = 2$, $\alpha = 0.1$ $\delta = 0.1/n$, $\sharp$Posterior $= \sharp$Langevin $= 15n$, $\sharp$MonteCarlo $= 50$

n	100	200	300
$\Delta(\alpha)$	4.368	4.417	**4.434**
TIC($\kappa = 0$)	92.92 $\pm$ 11.81	190.8 $\pm$ 16.99	289.8 $\pm$ 31.01
TIC($\kappa = 0.1$)	91.97 $\pm$ 11.71	134.3 $\pm$ 11.87	167.1 $\pm$ 17.84
FV(α)	4.127 $\pm$ 0.611	4.18 $\pm$ 0.341	4.315 $\pm$ 0.508
LFV(α)	3.498 $\pm$ 0.713	3.926 $\pm$ 0.457	**4.1** $\pm$ 0.522

*TIC with generalized inverse of information matrix (Thomas et al., 2020)

Numerical Experiments 2: Nonlinear NN

NN Architecture:

- $g_\theta(\mathbf{z}) = \langle \theta^{(2)}, \tanh(\theta^{(1)}\mathbf{z} + \theta^{(0)}) \rangle$ with $M = 50, 100, 150$ hidden units.
- $p := |\theta| = M(d+2)$.
- $\mu = g_{\theta_0}$ for some θ_0.
- NN is initialized by $\theta \sim N(\theta_0, 0.01\mathbf{I}_p)$.

Langevin FV Setting

- $T \in \{250, 1000\}$, and $\delta = 10^{-5}$. We use the last $0.9T$ iterations to compute LFV.

Generalization Gap

$$\tilde{\Delta} := \mathbb{E}_{\mathbf{y}^*}\left(\frac{1}{n}\sum_{i=1}^{n}\{y_i^* - g_{\hat{\theta}}(\mathbf{z}_i)\}^2\right) - \frac{1}{n}\sum_{i=1}^{n}\{y_i - g_{\hat{\theta}}(\mathbf{z}_i)\}^2$$

is computed over 50 times experiments.

Numerical Experiments 2: Nonlinear NN

Table: The generalization gap and LFV for the neural network model with $n = 1000$ and $T = 1000$. LFV values for the overparameterized regime (i.e., $p = M(d+2) > n$) are gray-colored.

	$M = 50$		$M = 100$		$M = 150$	
	LFV	$\tilde{\Delta}$	LFV	$\tilde{\Delta}$	LFV	$\tilde{\Delta}$
$d = 5$	6.43 ± 0.96	4.26	6.49 ± 0.52	4.41	7.30 ± 0.80	4.60
$d = 10$	11.03 ± 1.28	8.75	12.91 ± 1.54	9.18	13.56 ± 1.14	9.64
$d = 15$	16.78 ± 1.39	17.64	18.93 ± 1.57	18.60	20.13 ± 2.07	19.46

Summary

(1) We proved that $|\mathbb{E}_{\mathbf{y}}[\mathrm{FV}(\alpha)] - \Delta(\alpha)| \to^{p} 0$, for $p \geq n$.

(2) We proposed a Langevin FV, which is

- simply computed with GD-based packages (e.g., PyTorch),
- no needed to compute $p \times p$ information matrix (unlike TIC).

(3) We demonstrated Langevin FV in numerical experiments.

Contact Info: A. Okuno (*okuno@ism.ac.jp*)

- Another poster of ours (@IASC-ARS): "Algebraic Approach to Ridge-Regularized Mean Squared Error Minimization in Minimal ReLU Neural Network"
  - We enumerated all the local solutions for minimal ReLU NN (using symbolic computation).

References I

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. (2018). Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*.

Eaton, M. L. (1989). Group invariance applications in statistics. In *Regional conference series in Probability and Statistics*, pages i–133. JSTOR.

Karakida, R., Akaho, S., and Amari, S.-i. (2019). Universal statistics of fisher information in deep neural networks: Mean field approach. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1032–1041. PMLR.

References II

Thomas, V., Pedregosa, F., Merriënboer, B., Manzagol, P.-A., Bengio, Y., and Le Roux, N. (2020). On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 3503–3513. PMLR.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(116):3571–3594.

Watanabe, S. (2018). *Mathematical theory of Bayesian statistics*. CRC Press.