A generalization gap estimation for overparameterized models via Langevin functional variance

Akifumi Okuno\(^1\),\(^2\) and Keisuke Yano\(^1\)

\(^1\)The Institute of Statistical Mathematics
\(^2\)RIKEN Center for Advanced Intelligence Project

Abstract

This paper discusses estimating the generalization gap, a difference between a generalization gap and an empirical error, for overparameterized models (e.g., neural networks). We first show that a functional variance, a key concept in defining a widely-applicable information criterion, characterizes the generalization gap even in overparameterized settings, where a conventional theory cannot be applied. We next propose a computationally efficient approximation of the functional variance, a Langevin approximation of the functional variance (Langevin FV). This method leverages the 1st-order but not the 2nd-order gradient of the squared loss function; so, it can be computed efficiently and implemented consistently with gradient-based optimization algorithms. We demonstrate the Langevin FV numerically in estimating generalization gaps of overparameterized linear regression and non-linear neural network models.

Keywords: overparametrization, generalization, functional variance, Langevin approximation.

1 Introduction

The great success of deep neural networks (LeCun et al., 2015; Goodfellow et al., 2016) has been reported in many applied fields such as natural language processing, image processing, and natural science. This is now altering our understanding of a classical discipline in data science, the bias-variance dilemma: the richness of statistical models reduces the modeling bias but suffers from fitting spurious patterns and leading to poor prediction accuracy. However, the modern practice of deep neural networks suggests that it accommodates exactly fitting training data (known as interpolation) by an overabundance of parameters of models (known as over-parameterization) with accurately predicting test data (Chiyuan et al., 2017; Ruslan, 2017; Belkin, 2018). Several such surprising phenomena have been reported; the double descent phenomenon (Belkin, 2018; Mario et al., 2019, 2020; Preetum et al., 2020), and the multiple descent phenomenon (Adlam and Pennington, 2020; d’Ascoli et al., 2020). So, we are stuck in the situation where we need to rethink the bias-variance dilemma. Such mystery due to over-parameterization is now a central issue in modern data science.

Theoretical understanding of the generalization has been conducted for long decades. A useful insight into the generalization is derived by considering the worst-case performance. Vapnik and Chervonenkis (1971) show that the Vapnik–Chervonenkis (VC) dimension characterizes the worst-case generalization performance. Baum and Haussler (1989) analyze the VC dimension of neural networks to capture the generalization behaviour of neural networks. Another interesting approach to understanding the generalization is delivered by analyzing the deviance of the leave-one-out cross validation error relative to the generalization error. Devroye and Wagner (1979) give its upper bound using what is now called hypothesis stability (Kearns and Ron, 1999). Theoretical investigation in the generalization in over-parameterized models has been rapidly studied, and it is elucidated that over-parameterization may help reduce the generalization (Hastie et al., 2019; Nakada and Imaizumi, 2021; Adlam and Pennington, 2020).

Measuring the generalization gap, which is the gap between empirical and generalization errors, in a data-driven way helps us understand what is actually happening. Such tools have been developed in the literature of information criteria or optimism estimation. The well-known Akaike Information Criterion (AIC;
estimates the generalization log-loss of plug-in predictive density based on the maximum likelihood estimator. Takeuchi Information Criterion (TIC; Takeuchi, 1976) modifies AIC so as to deal with the model misspecification. Regularization Information Criterion (RIC; Shibata, 1989) extends TIC so as to accommodate maximum penalized likelihood estimation. Moody (1992) and Murata et al. (1994) generalize RIC to an arbitrary loss and advocate its use in non-linear systems such as neural networks. Mallows’ $C_p$ (Mallows, 1973) and Stein’s unbiased risk estimates (SURE; Stein, 1981) offer an elegant estimation scheme of the generalization gap using the covariance between a predictor and an outcome in Gaussian models with $\ell_2$ loss. Ramani et al. (2008) propose an efficient Monte Carlo sampling-based method (Monte-Carlo SURE) to estimate the covariance. Recently, new methodologies associated with Bayesian learning have been developed. The Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) and Widely-Applicable Information Criterion (WAIC; Watanabe, 2010) offer computationally-efficient devices that estimate the generalization error of Bayesian learning. In particular, Watanabe (2010) shows that in a statistical model with fixed dimension, the generalization gap of Bayesian learning is characterized by the posterior variance of log-likelihood, which is termed by the functional variance.

However, relatively scarce are studies of data-driven measurement of the generalization gap in the overparameterized regime. Gao and Jojic (2016) extend Monte-Carlo SURE to analyze the generalization gap in deep neural networks. Thomas et al. (2020) modify TIC for the same purpose. These studies empirically investigate the gaps in famous datasets such as MNIST and CIFAR-10, and successfully provide useful tools to understand the generalization gap in the over-parameterized regime. However, these approaches lack theoretical guarantees in overparameterized models, although both have theories in statistical models with fixed-dimension. So, the theoretical applicability remains unclear. Also, the computational costs of these approaches in both memory and speed are not low because the former requires training several times, and the latter needs the second-order gradient (Hessian) of the loss function.

In this paper, we focus on yet another tool to measure the generalization gap, the functional variance (FV) developed in Watanabe (2010) and provide the following contributions to the literature:

- **Theoretical applicability**: We prove that FV is asymptotically unbiased to the generalization gap of Bayesian learning for overparameterized linear regression models.

- **Computational efficiency**: We propose a computationally-efficient approximation of FV, Langevin FV (LFV), leveraging only the 1st-order gradient of the loss function.

Our theory employs an overparameterized linear regression model (Hastie et al., 2019; Belkin et al., 2020; Bartlett et al., 2020). It is, of course, a regression model with too many covariates and, what is more important, can be treated as a linear approximation of non-linear overparameterized models. This linearization plays a pivotal role in the theory of neural tangent kernel (Jacot et al., 2018; Arora et al., 2019). Although FV is defined by using the full posterior covariance, the full Bayesian inference in the overparameterized regime is often prohibited due to the computational burden. Then, inspired by the stochastic variational inference (SVI; Sato and Nakagawa, 2014; Mandt et al., 2017), we employ the Langevin dynamics (c.f., Risken, 1996; Cheng et al., 2018) of which the stationary distribution approximates the posterior distribution. This approach has several merits from the computational perspective; for example, the implementation is easy and consistent with the gradient-based optimization such as stochastic gradient descent, which is a de-facto standard in the deep neural network applications.

The rest of the paper is organized as follows. Section 2 provides the our theoretical results on FV, Section 3 proposes Langevin FV, which is an efficient implementation of FV using Langevin dynamics, Section 4 demonstrates numerical experiments, Section 5 provides proofs of main theorems, and Section 6 concludes this paper.

### 1.1 Symbols

Let $n, p \in \mathbb{N}$, and let $\mathbf{A} = (a_{ij}), \mathbf{B} = (b_{ij}) \in \mathbb{R}^{n \times p}$, $[n]$ denotes a set $\{1, 2, \ldots, n\}$, $\mathbf{A} \odot \mathbf{B} = (a_{ij} b_{ij}) \in \mathbb{R}^{n \times p}$ denotes the Hadamard element-wise product and $\mathbf{A}^{\otimes 2} := \mathbf{A}\mathbf{A}^\top \in \mathbb{R}^{n \times n}$. If $n = p$, $\text{tr} [\mathbf{A}] := \sum_{i=1}^{n} a_{ii}$. For any vector $\mathbf{p} = (\beta_1, \beta_2, \ldots, \beta_p) \in \mathbb{R}^p$, $\| \mathbf{p} \|_2 = (\beta_1^2 + \beta_2^2 + \cdots + \beta_p^2)^{1/2}$ and $\| \mathbf{p} \|_\infty = \max_{i \in [p]} | \beta_i |$. $\Theta (n) := \{ \mathbf{U} \in \mathbb{R}^{n \times n} \mid \mathbf{U}^\top \mathbf{U} = \mathbf{I}_n \}$ denotes a set of $n \times n$ orthogonal matrices and $S^{n-1} := \{ \mathbf{u} \in \mathbb{R}^n \mid \| \mathbf{u} \|_2 = 1 \}$ denotes a unit sphere surface. $\to^p$, $\to^d$ denote convergences in probability and in distribution, respectively.
2 Theoretical results for the functional variance

In this section, we present theoretical results for the functional variance in overparameterized linear regression models.

2.1 Problem setting

We begin with introducing the problem setting for the theory. We consider a linear regression model

\[ y = X\beta_0 + \epsilon, \]

where \( y = (y_1, \ldots, y_n) \in \mathbb{R}^n \) is a vector of observed outcomes, \( X = (x_1^\top, x_2^\top, \ldots, x_n^\top) \in \mathbb{R}^{n \times p} \) is an \( n \times p \) design matrix, \( \beta_0 \in \mathbb{R}^p \) is an unknown coefficient vector, and \( \epsilon = (\epsilon_1, \ldots, \epsilon_n) \) is a vector of i.i.d. error terms with mean zero and variance \( \sigma_0^2 \). Our interest is the overparameterized situation where the number of regressors \( p \) is larger than the sample size \( n \):

\[ n \leq p. \]

As we mentioned in the introduction, the overparameterized linear regression model can simulate a linear approximation of non-linear overparameterized models.

We take the quasi-Bayesian approach on the vector \( \hat{\beta} \). Working under the quasi-likelihood with a Gaussian distribution \( f(y \mid x_i, \hat{\beta}) \) with mean \( x_i^\top \hat{\beta} \) and variance \( \sigma_0^2 + \sigma_0^2 \) on \( \epsilon \) and a Gaussian prior \( \mathcal{N}_p(0, (\sigma_0^2/(an))I_p) \), we obtain the quasi-posterior distribution \( \beta \):

\[ \Pi_n(d\beta \mid y, X) = (2\pi)^{-p/2}(\det Q_n)^{-1/2}e^{-(\hat{\beta} - \hat{\beta}_n)Q_n^{-1}(\hat{\beta} - \hat{\beta}_n)/2}d\beta, \]

where \( \hat{\beta}_n \) is the maximum a posterior estimate

\[ \hat{\beta}_n := (n^{-1}X^\top X + \sigma_0^2 I_p)^{-1}n^{-1}X^\top y = \arg\min_{\beta \in \mathbb{R}^p} \{\ell_n(\beta) := n^{-1}\|y - X\beta\|_2^2 + \alpha\|\beta\|_2^2\} \]

and \( Q_n \) is the matrix

\[ Q_n := n^{-1}\sigma_0^2(n^{-1}X^\top X + \sigma_0^2 I_p)^{-1}. \]

We will analyze the Gibbs generalization gap

\[ \Delta(\alpha) := \frac{1}{2\sigma_0^2} \left\{ \mathbb{E}_y y^* \left[ \mathbb{E}_{\hat{\beta}} \left[ \|y^* - X\beta\|_2^2 \right] \right] - \mathbb{E}_y \left[ \mathbb{E}_{\hat{\beta}} \left[ \|y - X\beta\|_2^2 \right] \right] \right\}, \]

where \( y^* \) is an independent copy of \( y \mid X \), \( \mathbb{E}_{\hat{\beta}}[\cdot] \) is the expectation with respect to the quasi-posterior distribution \( 1 \). The Gibbs generalization gap considers one stochastic sample from the quasi-posterior distribution for the estimation. To estimate the Gibbs generalization gap from the current observations, we focus on the functional variance

\[ \text{FV}(\alpha) := \sum_{i=1}^n \mathbb{V}_{\hat{\beta}}[\log f(y_i \mid x_i, \hat{\beta})], \]

where \( y^* \) is an independent copy of \( y \mid X \), \( \mathbb{E}_{\hat{\beta}}[\cdot] \) is the expectation with respect to the quasi-posterior distribution \( 1 \), and \( \mathbb{V}_{\hat{\beta}} \) is the variance with respect to to the quasi-posterior distribution \( 1 \).

2.2 Asymptotic unbiasedness of the functional variance

Here we present our theoretical findings on the functional variance in the overparameterized regime. The following theorem shows that the functional variance \( \text{FV}(\alpha) \) is an asymptotically unbiased estimator of the Gibbs generalization gap \( \Delta(\alpha) \), for the overparameterized linear regression with Gaussian covariates. The proof is given in Section 5.
**Theorem 1.** Let \( n, p = p(n) \in \mathbb{N} \) with \( n \leq p \), \( \Sigma_n \) be a \( p \times p \) non-zero and non-negative definite matrix. Let \( x_1, x_2, \ldots, x_n \) be i.i.d. random vectors from Gaussian distribution with mean zero and covariance matrix \( \Sigma_n \). There exists an absolute constant \( C \) such that we have, for any \( \varepsilon, \alpha > 0 \),

\[
P \left( \left| \mathbb{E}_y [\text{FV}(\alpha)] - \Delta(\alpha) \right| > \varepsilon \right) \leq C \left( \xi^2 + \frac{1}{\alpha^2 n \varepsilon} + \frac{1}{\alpha^4 n^2 \varepsilon^2} + \frac{1}{n} \right),
\]

where \( \xi := \text{tr}(\Sigma_n) \) and \( b := p^{1/2} \| \beta_0 \|_\infty \). Further, under the conditions (C1) \( \xi < \infty \) and (C2) \( b < \infty \), we have \( |\mathbb{E}_y [\text{FV}(\alpha)] - \Delta(\alpha)| = O_p(a_n/\sqrt{n}) \) with arbitrary slowly increasing sequence \( a_n \).

**Theorem 1** implies that the FV successfully estimates the Gibbs generalization gap in the overparameterized regime, which supports the use of FV even in the overparameterized regime. Interestingly, its rate of convergence \( (a_n/\sqrt{n}) \) is dimension-free, that is, irrelevant of the dimension \( p \). Theorem 1 also elucidates the role of the trace of \( \Sigma_n \). As the trace gets larger, the decay of the difference becomes slower. Further, the result does not restrict the true distribution of the additive error to the Gaussian distribution; in contrast, some classical theories such as SURE require the error term to be Gaussian or related distributions.

Let us mention the conditions in the theorem. Condition (C1) indicates the trace boundedness of the Fisher information matrix \( \mathbf{X}^\top \mathbf{X} / n \) (divided by \( n \)). Both shallow and deep neural network models satisfy this condition in general settings (e.g., Karakida et al., 2019). Condition (C2) together with Condition (C1) indicates that \( \mathbf{u} = \mathbb{E}_y [y] = \mathbf{X} \hat{\beta}_0 \) satisfies \( n^{-1} \| \mathbf{u} \|_2^2 \leq n^{-1} \| \mathbf{X} \|_2^2 \| \beta_0 \|_2^2 \leq (n^{-1} \sum_{i=1}^n \mathbf{s}_{n,i}^2) (\sum_{i=1}^p \beta_0_i^2) = O_p(1) \), which implies the average of entries in the outcome expectation \( \mu \) is \( O_p(1) \).

The main ingredient of the proof of Theorem 1 is the explicit identity of the difference between the generalization gap and FV. Its proof is presented in Appendix A.

**Lemma 2.** We have

\[
\Delta(\alpha) = \text{tr}(H_\alpha) \quad \text{and} \quad \mathbb{E}_y [\text{FV}(\alpha)] - \Delta(\alpha) = -\frac{3}{2} \text{tr}(H_\alpha \circ H_\alpha) + \frac{1}{\sigma_0} \text{tr}(H_\alpha \circ ((I - H_\alpha)(X \beta_\rho)))^{\#2},
\]

where \( H_\alpha \) is the hat matrix

\[
H_\alpha := n^{-1} X (n^{-1} X^\top X + \alpha I_p)^{-1} X^\top.
\]

**Lemma 2** highlights the role of the hat matrix \( H_\alpha \) in evaluating the generalization gap and the residual. Same as the fixed-dimension theory, the hat matrix controls the magnitude of the generalization gap. With singular values \( s_1 \geq s_2 \geq \cdots \geq s_n \geq 0 \) of the matrix \( X \), under the conditions in Theorem 1, we have

\[
\Delta(\alpha) = \text{tr}(H_\alpha) = \sum_{i=1}^n \frac{s_{n,i}^2}{s_{n,i}^2 / n + \alpha} \rightarrow \exists \Delta(\alpha) \in [0, \infty) \quad (n \to \infty)
\]
as \( \text{tr}(H_\alpha) \leq \alpha^2 n^{-1} \sum_{i=1}^n s_{n,i}^2 < \infty \) with probability approaching 1 as \( n \to \infty \). Consider a simple case that \( X \) has a fixed intrinsic dimension \( p_* \ll n \) with growing \( n, p \), namely \( s_{n,1} \geq s_{n,2} \geq \cdots \geq s_{n,p_*} > 0 = s_{n,p_*+1} = s_{n,p_*+2} = \cdots = s_{n,n} \). Then, FV approaches the fixed intrinsic dimension \( p_* \) (as \( \alpha \to 0 \)), which coincides with a degrees of freedom (Mallows, 1973; Ye, 1998) and AIC (Akaike, 1974) for linear regression equipped with \( p_* \) regressors.

**Remark 1** (Sample-wise and joint log-likelihoods). One may think the use of the posterior covariance of the joint log-likelihood

\[
\text{J-FV}(\alpha) := \nabla_\beta \log f(y | X, \hat{\beta})
\]

instead of the FV (5) using sample-wise log-likelihood. In this case, we have the following expression of the difference between J-FV and the generalization gap. The proof is in Appendix B.

**Proposition 1.** For any \( \alpha > 0 \), we have

\[
\mathbb{E}_y [\text{J-FV}(\alpha)] - \Delta(\alpha) = -\frac{3}{2} \text{tr}(H_\alpha^2) + \frac{1}{\sigma_0} \text{tr}(H_\alpha (I - H_\alpha)(X \beta_\rho))^{\#2}.
\]
This exhibits an interesting correspondence between FV and J-FV (see Lemma 2 and Proposition 1); replacing the Hadamard products in $E_p[FV(\alpha)]$ with simple matrix products yields $E_p[J-FV(\alpha)]$. As the simple matrix products in $E_p[J-FV(\alpha)]$ do not vanish as $n \to \infty$, J-FV is not an asymptotically unbiased estimator of the generalization gap.

**Remark 2** (Generalization of the theorem). Theorem 1 is further generalized to Theorem 3, which proves the convergence of FV under milder conditions. The proof is shown in Section 5.

**Theorem 3.** Let $n, p = p(n) \in \mathbb{N}$ with $n \leq p$ and suppose the setting in Section 2.1. Let $q_{n,1}, q_{n,2}, \ldots, q_{n,n}$ be marginal probability densities of the left-singular vectors $u_{n,1}, u_{n,2}, \ldots, u_{n,n} \in \mathbb{S}^{n-1}$ of a random matrix $X \in \mathbb{R}^{n \times p}$. Write

$$b := p^{1/2} \|eta_0\|_{\infty}, \quad \psi := \max_{u \in \mathbb{S}^{n-1}} q_{n,i}(u) \int_{\mathbb{S}^{n-1}} dv, \quad \eta := \frac{1}{n} \langle r|X^T X \rangle.$$

Then, there exists an absolute constant $C > 0$ such that we have, for any $\epsilon, \alpha > 0$,

$$\mathbb{P}(\|E_p(FV(\alpha)) - \Delta(\alpha)\| > \epsilon | X) \leq C \left( \frac{\psi \eta^2}{\alpha^2} \frac{1}{ne} + \frac{\psi \eta^4}{\alpha^4} \frac{1}{n \epsilon} + \frac{\psi \eta^4 b^4}{\alpha^2 \sigma_0^4} \frac{1}{n \epsilon^2} \right).$$

**Theorem 1** (Gaussian covariates) is covered as the case with $\psi = 1$ in Theorem 3; see Proposition 7.1 of Eaton (1989) which proves that the left-singular vectors of the Gaussian design matrix follow a uniform distribution over the sphere surface $\mathbb{S}^{n-1}$, i.e., $q_{n,i}(u) = 1(u \in \mathbb{S}^{n-1})/\int_{\mathbb{S}^{n-1}} dv$.

**Remark 3** (Definition of the generalization gap). Regarding the definition of the generalization gap $\Delta(\alpha)$, we may consider another design matrix $X^*$ for generating $y^*$. However, the covariate difference is less effective to the generalization gap (4), as we focus on the prediction of the conditional random variable $y|X$ but not the covariate $X$. For theoretical simplicity, we employed a single design matrix $X$ to generate both outcomes $y, y^*$.

## 3 Langevin Functional Variance

Whereas FV is theoretically attractive as proved in Theorem 3, there exist two computational difficulties:

1. **(D1)** generating samples from the quasi-posterior (1) requires $p \times p$ matrix $Q_\alpha$, which is computationally intensive for overparameterized models $n \ll p$, and
2. **(D2)** computing the quasi-posterior (1) is inconsistent with the gradient-based optimization approaches, such as stochastic gradient descent (SGD), that are often used in optimizing overparameterized models.

To resolve these difficulties (D1) and (D2), we propose an efficient approximation of FV using Langevin dynamics.

### 3.1 Langevin approximation of the quasi-posterior

A key idea of our algorithm is approximating the quasi-posterior (1) by a Langevin process: starting with an estimator $y^{(0)} := \hat{\beta}_a$, we stochastically update $y^{(t)}$ by

$$y^{(t+1)} = y^{(t)} - \frac{1}{4} \delta \kappa_n \frac{\partial \ell_\alpha(y^{(t)})}{\partial y} + \delta^{1/2} e^{(t)} \quad (t = 1, 2, \cdots),$$

where $\kappa_n := n/\sigma_0^2$ and $\delta > 0$ are user-specified parameters, $e^{(1)}, e^{(2)}, \cdots \sim N_p(0, I_p)$ denote i.i.d. normal random vectors and $\ell_\alpha$ is defined in (2). We here discuss why this approximation works. First, the Langevin process (9) is a discretization of Ornstein-Uhlenbeck process

$$d\tilde{y}_\tau = -\frac{1}{2} Q_\alpha^{-1}(\tilde{y}_\tau - \hat{\beta}_a) d\tau + de_\tau,$$
equipped with a Wiener process \( \{e_t\}_{t \geq 0} \), i.e., \( e_t - e_{t'} \sim N_p(0, (\tau - \tau') I_p) \) for any \( \tau > \tau' \geq 0 \). The distribution of \( \tilde{\mathbf{y}}_t \) with the initialization \( \tilde{\mathbf{y}}_0 := \tilde{\mathbf{b}}_a \) coincides with the quasi-posterior

\[
\mathcal{D}(\tilde{\mathbf{y}}_t) = N_p(\tilde{\mathbf{b}}_a, Q_a) \quad (\tau > 0);
\]

(10)

See, e.g., Risken (1996) page 156 for the distribution (10). Next, Theorem 2 of Cheng et al. (2018) evaluates 1-Wasserstein distance (a.k.a. Earth-mover’s distance) between distributions of OU and Langevin processes as

\[
W_1(\mathcal{D}(\tilde{\mathbf{y}}_t), \mathcal{D}(\mathbf{y}^{(t)})) \leq \epsilon
\]

(11)

for some \( \delta = O(e^2 / \rho) \) and \( t = p/\epsilon^2 \) for sufficiently small \( \epsilon > 0 \); see Proposition 2 in Supplement C for details. So, the relations (10) and (11) imply that the Langevin process distribution \( \mathcal{D}(\mathbf{y}^{(t)}) \) approximates the quasi-posterior \( N_p(\tilde{\mathbf{b}}_a, Q_a) \).

This Langevin approximation resolves the difficulties (D1) and (D2) as

(S1) the Langevin process (9) is computed with only the 1st order gradient of the squared loss function \( \ell_a(\mathbf{y}) \) defined in (2), meaning that the large matrix \( Q_a \) is not explicitly computed, and

(S2) the Langevin process (9) is in the form of gradient descent up to the normal noise \( \delta^{1/2} e^{(t)} \); this process can be implemented consistently with gradient-based algorithms, which are often used for optimizing overparameterized models.

**Remark 4** (Comparison with the stochastic natural gradient descent). The Langevin approximation is related to stochastic variational inference (SVI; Sato and Nakagawa, 2014; Mandt et al., 2017), which proves that stochastic natural gradient descent (SNGD; Amari, 1998; Amari and Douglas, 1998) with a specific constant stepsize approximates the quasi-posterior by assuming that the stochastic noise in SNGD approximates the normal noise \( e^{(t)} \) in Langevin dynamics (9). However, this normal approximation is hardly satisfied for overparameterized setting \( n \leq p \), where only a small \( p \) is considered in most SVI references. Thus we employ the simple Langevin dynamics (9), which does not rely on the normal approximation and is practically more stable than SVI in the overparameterization setting. Yet, we can replace the Langevin process (9) with other stochastic processes, as long as they approximate the quasi-posterior (1).

### 3.2 Langevin functional variance

Using the Langevin approximation (9), we propose a Langevin functional variance (LFV): For a time step \( T \in \mathbb{N} \), let \( \{\mathbf{y}^{(t)}\}_{t=1,2,\ldots,T} \) be the samples from the Langevin process. Let \( \hat{V}_T[\cdot] \) denotes an empirical variance with respect to \( \{\mathbf{y}^{(t)}\}_{t=1,2,\ldots,T} \). Then, we define a Langevin functional variance (LFV) as

\[
\text{LFV}(\alpha) := \frac{1}{n} \sum_{i=1}^n \hat{V}_T[\log f(y_i | x_i, \mathbf{y})] = \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T \left\{ \frac{1}{2\sigma_a^2} (y_i - \mu_i^{(t)})^2 - \frac{1}{T} \sum_{t'=1}^T (y_i - \mu_i^{(t')})^2 \right\}.
\]

(12)

As the Langevin process approaches the quasi-posterior, LFV is expected to approximate FV with sufficiently large \( T \in \mathbb{N} \).

Whereas the distribution approximation of high-dimensional vector \( \mathbf{y} \in \mathbb{R}^p \) faces the curse of dimensionality, FV is a statistic of 1-dimension; FV and LFV approximate the generalization gap \( \Delta(\alpha) \) with a reasonable number of samples \( T \in \mathbb{N} \), even if the dimension \( p \) is relatively large. See numerical experiments in Section 4.1.

### 3.3 Application to non-linear models

Herein, we apply LFV (12) to a nonlinear overparameterized model \( g_\theta(z_i) \) equipped with a parameter vector \( \theta \) and an input vector \( z_i \). The gradient of the overparameterized model \( g_\theta \) is compatible with the gradient of its linear approximation, where this linear approximation is represented by a linear regression model in this paper; we replace \( \ell_a(\mathbf{b}) \) in the Langevin process (9) with

\[
\rho_a(\mathbf{b}) := \frac{1}{n} \sum_{i=1}^n (y_i - g_\theta(z_i))^2 + \alpha \|\mathbf{b}\|_2^2.
\]
and the Langevin sequence \( \{y^{(t)}_i\}_{t=1,2,...} \) yields LFV for the nonlinear model \( g_\theta \) by substituting \( \mu^{(t)}_j := g_{y^{(t)}}(z_t) \) to (12). See, e.g., Dalalyan (2017) for nonlinear extensions of Langevin approximation.

### 3.4 Comparison to the existing methods

We here compare FV and LFV to the existing method for estimating the generalization gap using Takeuchi Information Criterion (TIC; Takeuchi, 1976). TIC estimates the generalization gap of regular prediction models by using a quantity

\[
\text{tr}\left\{G \hat{F}^{-1}\right\}
\]

(13)
equipped with \( p \times p \) matrices

\[
G := \sum_{i=1}^n \frac{\partial \log f(y_i | x_i, \beta)}{\partial \beta} \frac{\partial \log f(y_i | x_i, \beta)}{\partial \beta^\top} \quad \text{and} \quad F := \sum_{i=1}^n \frac{\partial^2 \log f(y_i | x_i, \beta)}{\partial \beta \partial \beta^\top}.
\]

However, TIC cannot be applied to singular prediction models such as overparameterized models because the Fisher information matrix \( \hat{F} \) degenerates, whereby the inverse of \( \hat{F} \) does not exist. To overcome this limitation of TIC, Thomas et al. (2020) replaces the inverse matrix \( \hat{F}^{-1} \) in (13) with the \( \kappa \)-generalized inverse matrix \( \hat{F}_\kappa^+ \). \( \hat{F}_\kappa^+ \) is defined as \( U \Sigma^+ V \), where \( U \) and \( V \) are matrices of the left- and right- singular vectors of \( \hat{F} \), respectively, and \( \Sigma^+ \) is the diagonal matrix of which the \( j \)-th diagonal component is

\[
\lambda_j(\hat{F}_\kappa^+) := \begin{cases} 
1/\lambda_j(\hat{F}) & (\lambda_j(\hat{F}) > \kappa) \\
0 & (\lambda_j(\hat{F}) \leq \kappa)
\end{cases}, \quad (j = 1, 2, \ldots, p)
\]

with \( \{\lambda_j(\hat{F}) : j = 1, \ldots, p\} \) singular values of \( \hat{F} \). This TIC modification of Thomas et al. (2020) is numerically examined in Section 4.1; it rather estimates the number of nonzero eigenvalues in \( \hat{F} = n^{-1}X^\top X \), i.e., the number of nonzero singular values of \( X \), but not the generalization gap \( \Delta(\alpha) \).

Instead of modifying TIC, we can employ the regularized information criterion (RIC) proposed by Shibata (1976). RIC replaces the Fisher information matrix \( \hat{F} \) in TIC with the regularized inverse matrix \( (\hat{F} + \alpha I)^{-1} \). Moody (1992) and Murata et al. (1994) further generalize RIC to arbitrary loss functions and demonstrate its use for shallow neural network models. RIC have much in common with FV; especially, RIC is also an asymptotically unbiased estimator of the Gibbs generalization gap \( \Delta(\alpha) \) for overparameterized linear regression, and numerically, RIC behaves similarly to FV (and LFV) for the experiments in Section 4.1. However, RIC still requires the inverse matrix of \( p \times p \) matrix \( \hat{F} + \alpha I \), and its computational cost is intensive for overparameterized setting \( p \geq n \).

### 4 Numerical Experiments

This section evaluates the Langevin FV (12) by using synthetic datasets of both linear and nonlinear models.

#### 4.1 Experiments on the convergence of FV and LFV

We first evaluate the Langevin FV (12) in the comparison with some baselines by synthetic dataset experiments of linear overparameterized models.

The set-up of numerical experiments is summarized as follows:

- **Synthetic data generation**: for \( p = 2n \) with \( n = 100, 200, 300, 400 \), orthogonal matrices \( U \in \mathbb{R}^{n \times n} \) and \( V \in \mathbb{R}^{P \times n} \) are i.i.d. from the uniform distribution over the set of orthogonal matrices satisfying \( U^\top U = V^\top V = I_n \), respectively. Entries of \( \beta_0 \in \mathbb{R}^P \) are i.i.d. from \( N(0, 1/p) \). With user-specified singular values \( \{s_i\}_{i=1}^n \), the design matrix \( X := USV^\top \) is computed. \( y \) is generated 50 times from a normal distribution \( \mathcal{N}_0(X\beta_0, I_n) \), i.e., \( \sigma_0^2 = 1 \).

- **Evaluation**: We evaluate the following baselines and LFV with 50 times experiments. Throughout these experiments, we employ \( \alpha = 0.1 \) for ridge regularization (2).
(a) **TIC penalty** (Takeuchi, 1976) is extended to overparameterization setting (Thomas et al., 2020), by replacing the inverse of Fisher information matrix \( \hat{F} \) in TIC with the generalized inverse matrix \( \hat{F}_+ \). See Appendix 3.4 for the definition of \( \text{TIC}(\alpha) := \text{tr}(\hat{G}_+\hat{F}_+) \).

(b) **FV** is empirically computed with \( T = 15n \) samples of \( \beta \) generated from the quasi-posterior (1). Its theoretical expectation shown in Lemma 2 is also computed.

(c) **LFV** is computed with \( \delta = 1/(10n) \) and \( T = 15n \) Langevin samples \( \gamma \) (see eq. (9)).

We consider three different types of singular values: (i) \( X \) has a fixed intrinsic dimension \( d_s = 5 \), i.e., \( s_1 = \cdots = s_5 = n^{1/2} \) and \( s_6 = s_7 = \cdots = s_d = 0 \), (ii) \( s_i = n^{1/2}i^{-1} \), and (iii) \( s_i = n^{1/2}i^{-1/2} \). Theorem 3 proves the convergence of the FV in the settings (i) and (ii) as \( \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} s_i^2 < \infty \) in (i) and (ii), whereas it does not proves the convergence in the setting (iii) as \( \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} s_i^2 \) is not bounded in (iii).

Results are shown in the following Table 1–3. Overall, FV and LFV can estimate the generalization gap \( \Delta(\alpha) \) \( (p = 2n, n \to \infty) \) well compared to TIC, and their biases are not drastically different. TIC is totally inaccurate for the settings (ii) and (iii); together with the result of (i), TIC rather estimates the number of nonzero eigenvalues in the Fisher information matrix \( \hat{F} \), which is generally different from the generalization gap \( \Delta(\alpha) \).

### Table 1: The generalization gap \( \Delta(\alpha) \) and its estimates (TIC, FV, LFV) with the standard deviations in setting (i) where \( s_i = n^{1/2}i^{-1} \) for \( i = 1, 2, \ldots, 10 \) and 0 otherwise. The estimate closest to the generalization gap is underlined in each column.

| \( n \) | 100 | 200 | 300 | 400 |
| --- | --- | --- | --- | --- |
| \( \Delta(\alpha) \) | 9.091 | 9.091 | 9.091 | 9.091 |
| TIC(\( \kappa = 0 \)) | 8.770 ± 1.412 | 9.473 ± 1.035 | 9.466 ± 0.885 | 9.908 ± 0.787 |
| TIC(\( \kappa = 0.1 \)) | 8.770 ± 1.412 | 9.473 ± 1.035 | 9.466 ± 0.885 | 9.908 ± 0.787 |
| FV(\( \alpha \)) | 8.458 ± 1.286 | 8.845 ± 0.941 | 8.789 ± 0.802 | 9.121 ± 0.729 |
| \( E_y[FV(\alpha)] \) | 8.533 | 8.804 | 8.894 | 8.948 |
| LFV(\( \alpha \)) | 8.477 ± 1.368 | 8.966 ± 0.9943 | 8.917 ± 0.792 | 9.328 ± 0.796 |

### Table 2: The generalization gap \( \Delta(\alpha) \) and its estimates (TIC, FV, LFV) with the standard deviations in setting (ii) where \( s_i = n^{1/2}i^{-1} \). The estimate closest to the generalization gap is underlined in each column.

| \( n \) | 100 | 200 | 300 | 400 |
| --- | --- | --- | --- | --- |
| \( \Delta(\alpha) \) | 4.368 | 4.417 | 4.434 | 4.442 |
| TIC(\( \kappa = 0 \)) | 92.92 ± 11.81 | 190.8 ± 16.99 | 289.8 ± 31.01 | 392.5 ± 25.65 |
| TIC(\( \kappa = 0.1 \)) | 91.97 ± 11.71 | 134.3 ± 11.87 | 167.1 ± 17.84 | 195.3 ± 12.58 |
| FV(\( \alpha \)) | 4.127 ± 0.611 | 4.180 ± 0.341 | 4.315 ± 0.508 | 4.402 ± 0.286 |
| \( E_y[FV(\alpha)] \) | 4.147 | 4.304 | 4.358 | 4.386 |
| LFV(\( \alpha \)) | 3.498 ± 0.713 | 3.926 ± 0.457 | 4.1 ± 0.522 | 4.292 ± 0.328 |

### 4.2 Experiments on LFV for non-linear models

In this section, we evaluate the Langevin FV for non-linear neural networks by synthetic dataset experiments. See Section 3.3 for application of LFV to non-linear models.

The set-up is summarized as follows:

- **Synthetic data**: Let \( n = 1000, \sigma^2 = 1 \). For \( i = 1, 2, \ldots, n \), we randomly generate \( z_i \overset{\text{iid}}{\sim} N_d(0, I_d) \), \( \mu_i := \mu(z_i), y_i := \mu_i + \varepsilon_i \) and \( \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \) with a function \( \mu(z) := 3 \tanh(\langle z, 1 \rangle / 2) \).
Table 3: The generalization gap $\Delta(a)$ and its estimates (TIC, FV, LFV) with the standard deviations in setting (iii) where $s_i = n^{1/2} \eta^{-1/2}$. The estimate closest to the generalization gap is underlined in each column.

| $n$  | 100  | 200  | 300  | 400  |
|------|------|------|------|------|
| $\Delta(a)$ | 23.53 | 29.98 | 33.86 | 36.66 |
| TIC ($\kappa = 0$) | 62.38 ± 10.14 | 154.3 ± 16.29 | 242.0 ± 22.27 | 333.5 ± 25.92 |
| TIC ($\kappa = 0.1$) | 62.38 ± 10.14 | 154.3 ± 16.29 | 242.0 ± 22.27 | 333.5 ± 25.92 |
| FV($\alpha$) | 17.38 ± 2.397 | 25.40 ± 2.454 | 29.22 ± 2.544 | 32.25 ± 2.394 |
| $E_y[FV(\alpha)]$ | 17.34 ± 2.301 | 24.69  | 29.26  | 32.57  |
| LFV($\alpha$) | 13.95 ± 2.301 | 22.73 ± 2.554 | 27.22 ± 2.575 | 30.54 ± 2.351 |

- **Neural network**: We employ a fully-connected one-hidden-layer neural network

$$g_\theta(z) := (\theta^{(2)}, \tanh(\theta^{(1)} z + \theta^{(0)}))$$

with $M \in \{50, 100, 150\}$ hidden units, where $\theta = (\theta^{(0)}, \theta^{(1)}, \theta^{(2)}) \in \mathbb{R}^M \times \mathbb{R}^{M \times d} \times \mathbb{R}^M$ is a parameter vector (and so the number of parameter is $p = M(d + 2)$) and $\tanh$ applies hyperbolic tangent function $\tanh(z) := (\exp(z) - \exp(-z))/(\exp(z) + \exp(-z))$ entry-wise. Let $\theta_0$ be a parameter satisfying $g_{\theta_0} = \mu$: for each experiment, we initialize the parameter $\theta$ by the element-wise independent normal distribution whose mean is $\theta_0$ with the element-wise variance 0.01, and update $\theta$ by full-batch gradient descent with learning rate 0.1, ridge-regularization coefficient $\alpha = 10^{-3}$ and 100 iterations.

- **Langevin FV**: For each setting $(d, M) \in \{5, 10, 15\} \times \{50, 100, 150\}$, we take average of LFV over 20 times experiments. In each experiment, we randomly generate $\{(z_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, train the NN $g_\theta$, and compute $T \in \{250, 1000\}$ iterations of Langevin process with $\delta = 10^{-3}$. We discard the first 0.1T iterations, and employ the remaining 0.9T iterations to compute Langevin FV.

- **Generalization gap $\tilde{\Delta}$**: For each setting $(d, M) \in \{5, 10, 15\} \times \{50, 100, 150\}$, we take average of the following generalization gap

$$\tilde{\Delta} := E_y \left( \frac{1}{n} \sum_{i=1}^n \{y_i^* - g_{\theta_{\Delta}}(z_i)\}^2 \right) - \frac{1}{n} \sum_{i=1}^n \{y_i - g_{\theta_{\Delta}}(z_i)\}^2$$

over 50 times experiments. In each experiment, we randomly generate $\{(z_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, train the NN $g_\theta$ and evaluate $\tilde{\Delta}$. We employ the ground-truth $\mu_i = \mu(z_i)$ and $\sigma^2 = 1$.

Table 4: The generalization gap and LFV for the neural network model with $n = 500$ and $T = 1000$. LFV values for the overparameterized regime (i.e., $p = M(d + 2) > n$) are gray-colored.

| $d$ | 5  | 10  | 15  |
|-----|----|-----|-----|
| $M = 50$ | 4.79 ± 0.67 | 9.10 ± 1.06 | 12.96 ± 1.06 |
| $M = 100$ | 4.80 ± 0.67 | 9.14 ± 1.06 | 12.98 ± 1.06 |
| $M = 150$ | 4.80 ± 0.67 | 10.78 ± 0.67 | 12.98 ± 0.67 |

Langevin FV (LFV) and generalization error $\tilde{\Delta}$ for non-linear neural network with $M$ hidden units are shown in the following Table 4–6. Therein, LFV values for overparameterized models (i.e., $p = M(d + 2) > n$) are colored gray. First, all tables show that LFV roughly estimates the generalization gap even for non-linear overparameterized neural network models. Second, Tables 5 and 6 suggest that LFV with fewer $T = 250$ iteration rather underestimates $\tilde{\Delta}$ compared to LFV with $T = 1000$ iteration. Third, Tables 4 and 5 suggest that the biases of LFV reduces as the sample size grows.
We then provide supporting lemmas and prove Theorem 1. Together with Theorem 4.

Lemma 3. Let \( S \) distributed over the unit sphere surface \( \mathbb{S}^{d-1} \), the Chebyshev’s inequality proves

\[
\mathbb{P}(v - E[v]) > \epsilon) \leq \frac{\epsilon^2}{\sigma^2}
\]

Then, it holds for an absolute constant \( C \).

We present supporting lemmas for the proof of Theorem 5.2. Supporting lemmas

Proof of Theorem 1

With the \( i \)-th diagonal entry \( \sigma^2_{n,i} \) of \( \Sigma_n \), write \( \bar{\mu}_n := \sum_{i=1}^n \sigma^2_i := (\text{tr}(\bar{\Sigma})) \) and \( \bar{\sigma}_n := 2\sum_{i=1}^n \sigma^2_i (\leq 2(\text{tr}(\Sigma_n))^2) \).

Then, it holds for \( \ell_i := \|x_i\|^2 \) and \( N(\mu_n, \bar{\sigma}_n) \) that \( \eta = n^{-1}\text{tr}(X^T X) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n x_{ij}^2 = n^{-1} \sum_{i=1}^n \ell_i \), whereby Chebyshev’s inequality proves

\[
\mathbb{P}(|\eta - \bar{\mu}_n| > \bar{\sigma}_n) \leq 1/n.
\]

Together with Theorem 3 and \( \psi = 1 \) for Gaussian covariates, this yields

\[
\mathbb{P}(|E_y[FV(\alpha)] - \Delta(\alpha)| > \epsilon) \leq \mathbb{P}(|E_y[FV(\alpha)] - \Delta(\alpha)| > \epsilon | \eta - \bar{\mu}_n - \bar{\sigma}_n) + \mathbb{P}(|\eta - \bar{\mu}_n| > \bar{\sigma}_n)
\]

\[
\leq C' \left( \frac{(\bar{\mu}_n + \bar{\sigma}_n)^2}{\epsilon^2} + \frac{(\bar{\mu}_n + \bar{\sigma}_n)^4}{\epsilon^4} + \frac{(\bar{\mu}_n + \bar{\sigma}_n)^6}{\epsilon^6} \right) + \frac{1}{n}
\]

for an absolute constant \( C' > 0 \). Considering the inequality \( \bar{\mu}_n + \bar{\sigma}_n \leq \text{tr}(\Sigma_n) + \sqrt{2} \text{tr}(\Sigma_n) \leq (1 + \sqrt{2}) \text{tr}(\Sigma_n) \), we conclude the assertion. \(\square\)

5.2 Supporting lemmas

We present supporting lemmas for the proof of Theorem 3. The first lemma gives the explicit form of the expectation of \((v^T Av)^2\). The second lemma provides the tail bound of the trace of Hadamard product.

Lemma 4. Let \( v \in \mathbb{R} \) and let \( A \in \mathbb{R}^{n \times n} \) be a diagonal matrix. For a random vector \( v \) which is uniformly distributed over the unit sphere surface \( \mathbb{S}^{n-1} \), we have

\[
\mathbb{E}[(v^T Av)^2] = \frac{2\text{tr}(A^2) + \text{tr}(A)^2}{n(n+2)}.
\]

Proof of Lemma 4. Let \( v_i \) be the \( i \)-th component of \( v \). Let

\[
\alpha_1 := \mathbb{E}[v_i^2 v_j^2] \quad \text{and} \quad \alpha_2 := \mathbb{E}[v_i^4].
\]
We begin with showing
\[
\alpha_1 = \mathbb{E}[v_1^2 v_2^2] = \frac{1}{n(n+2)}. \tag{15}
\]

Consider a random variable \( z = (z_1, z_2, \ldots, z_n) \in \mathbb{R}^n \) following a standard normal distribution \( N_n(0, I_n) \) and \( \tilde{v} := (\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_n) := z/\| z \|_2 \). Then, the random vector \( \tilde{v} \) follows the uniform distribution over the sphere surface \( S^{n-1} \) (see, e.g., Eaton (1989) Proposition 7.1). Since \( 1/\| z \|_2^2 = \int_0^\infty \exp(-t/\| z \|_2^2) \, dt \), we get
\[
\alpha_1 = \mathbb{E}[\tilde{v}_1^2 \tilde{v}_2^2] = \mathbb{E}[z_1^2 z_2^2/\| z \|_2^2] = \mathbb{E}\left[ z_1^2 z_2^2 \int_0^\infty \exp(-t/\| z \|_2^2) \, dt \right].
\]

Together with the identity \( \int_0^\infty \exp(-t/\| z \|_2^2) \, dt = \int_0^\infty \prod_{i=1}^n \exp(-t z_i^2) \, dt \), this yields
\[
\alpha_1 = \mathbb{E}\left[ z_1^2 z_2^2 \int_0^\infty \prod_{i=1}^n \exp(-t z_i^2) \, dt \right] = \int_0^\infty \mathbb{E}[z_1^2 \exp(-t z_1^2)] \mathbb{E}[z_2^2 \exp(-t z_2^2)] \prod_{i=3}^n \mathbb{E}[\exp(-t z_i^2)] \, dt.
\]

Using a function \( u(t) := \mathbb{E}[\exp(-t z_i^2)] = (1 + 2t)^{-1/2} \), we have
\[
\alpha_1 = \int_0^\infty u'(t)^2 u(t)^{n/2} \, dt = \int_0^\infty (1 + 2t)^{-n/2 - 2} \, dt,
\]
which yields
\[
\alpha_1 = \frac{1}{2n(1+2t)^{n/2}} - \frac{1}{2(n+2)(1+2t)^{n/2+1}} \bigg|_0^\infty = \frac{1}{n(n+2)}
\]
and thus we get (15).

Taking expectation to \( 1 = \| v \|_2^2 = \sum_{i,j} v_i^2 v_j^2 + \sum_i v_i^4 \) yields an equation \( 1 = n(n-1)\alpha_1 + n\alpha_2 \), indicating that \( \alpha_2 = 1/(n(n+2)) \). This, together with \( \alpha_1 = 1/(n(n+2)) \), gives
\[
\mathbb{E}[(v^\top Av)^2] = \sum_k a_k^2 \mathbb{E}[v_k^4] + \sum_{k \neq l} a_k a_l \mathbb{E}[v_k^2 v_l^2] = \alpha_2 \sum_k a_k^2 + \alpha_1 \sum_{k \neq l} a_k a_l = (\alpha_2 - \alpha_1) \sum_k a_k^2 + \alpha_1 \left( \sum_k a_k \right)^2 = \frac{2\text{tr}(A^2)}{n(n+2)} + \frac{(\text{tr}(A))^2}{n(n+2)},
\]
which concludes the proof.

**Lemma 5.** Let \( n \in \mathbb{N} \) and define a diagonal matrix \( R \in \mathbb{R}^{n \times n} \) whose diagonal entries are \( \{r_j\}_{j=1}^n \), and let \( U = (u_1, u_2, \ldots, u_n) \in \mathcal{O}(n) \) be a random orthogonal matrix where marginal density functions of the row vectors \( u_1, u_2, \ldots, u_n \) are \( q_{n,1}, q_{n,2}, \ldots, q_{n,n} \), respectively. Write
\[
\psi := \max_{u \in \mathbb{S}^{n-1}} q_{n,i}(u) \int_{\mathbb{S}^{n-1}} \, dv.
\]

Then, it holds for \( T = (t_{ij}) := URU^\top \) and \( \varepsilon > 0 \) that
\[
P(\text{tr}(T \circ T) > \varepsilon) < \frac{3\psi}{n\varepsilon^2}(\text{tr}(R))^2.
\]

**Proof of Lemma 5.** Write \( v_n := \int_{\mathbb{S}^{n-1}} \, dv. \) Since \( t_{ii} = u_i^\top R u_i \), we have
\[
\max_{i=1,2,\ldots,n} \mathbb{E}[t_{ii}^2] = \max_{i=1,2,\ldots,n} \mathbb{E}[(u_i^\top R u_i)^2] = \max_{i=1,2,\ldots,n} \int_{\mathbb{S}^{n-1}} (u_i^\top R u_i)^2 q_{n,i}(u_i) \, du_i.
\]
where the random vector $v$ follows a uniform distribution over the set of orthogonal matrices $\mathbb{S}^{n-1}$. Together with Lemma 4, this gives
\[
\max_{i=1,2,\ldots,n} \mathbb{E}[t_{ii}^2] = \frac{2\text{tr}(R^2) + (\text{tr}(R))^2}{n(n+2)} \leq \frac{\psi}{n^2}[2\text{tr}(R^2) + (\text{tr}(R))^2].
\]
Therefore, the Cauchy–Schwarz inequality for the trace gives
\[
\mathbb{E}\left[\sum_{i=1}^{n} t_{ii}^2 \right] \leq n \max_{i=1,2,\ldots,n} \mathbb{E}[t_{ii}^2] \leq \frac{\psi}{n}[2\text{tr}(R^2) + (\text{tr}(R))^2] \leq \frac{3\psi}{n}(\text{tr}(R))^2.
\]
Substituting the expectation bound (16) to Markov’s inequality $\mathbb{P}(\sum_{i=1}^{n} t_{ii}^2 > \epsilon) < \mathbb{E}[\sum_{i=1}^{n} t_{ii}^2]/\epsilon$ proves the assertion. \hfill \Box

### 5.3 Proof of Theorem 3

We first prepare some notations for the proof. We denote by $\mathbb{P}_n$ the conditional probability $\mathbb{P}(\cdot | X)$. The decomposition $X = U S V^\top$ denotes a singular-value decomposition, where $S$ is a diagonal matrix whose diagonal entries $s_{n,1} \geq s_{n,2} \geq \cdots \geq s_{n,n} \geq 0$ are singular values of $X$. So, we have $\eta = n^{-1}\text{tr}(X^\top X) = n^{-1}\sum s_{n,i}^2$. The matrix $H_a$ is the hat matrix in (6). Let $\delta_n^{(1)} := \text{tr}(H_a \circ H_a), \delta_n^{(2)} := \text{tr}(H_a \circ H_a^2)$ and $\delta_n^{(3)} := \text{tr}(H_a \circ (I_n - H_a) (X \beta_a)_n))^{\alpha^2}/\sigma_a^2$, respectively.

Considering Lemma 2, we only have to prove tail bounds for $\delta_n^{(1)}, \delta_n^{(2)},$ and $\delta_n^{(3)}$ via three steps.

**The first step: Tail bound for $\delta_n^{(1)}$.** From the singular value decomposition $X = U S V^\top$, we have
\[
H_a = U R^{(1)} U^\top,
\]
where $R^{(1)}$ is the diagonal matrix whose diagonal entries $r_{1}^{(1)}, r_{2}^{(1)}, \ldots, r_{n}^{(1)}$ are given by
\[
r_{i,n}^{(1)} := \left(\frac{s_{i,n}^2}{n}/\alpha \right) \
\leq \left(\frac{s_{i,n}^2}{n} + \alpha\right)/\alpha.
\]
Then, applying Lemma 5 to $H_a$, we get
\[
\mathbb{P}_n(\delta_n^{(1)} > \epsilon) \leq \frac{3\psi}{\alpha \epsilon}(\text{tr}(R^{(1)}))^2 \leq \frac{3\psi}{\alpha \epsilon \alpha^2} \epsilon^2,
\]
which completes the first step.

**The second step: Tail bound for $\delta_n^{(2)}$.** From the singular value decomposition $H_a = U R^{(1)} U^\top$, we have
\[
H_a^2 = U R^{(2)} U^\top,
\]
where $R^{(2)} = (R^{(1)})^2$ and so
\[
\text{tr}(R^{(2)}) = \sum_{i=1}^{n} r_{i,n}^{(2)} \leq \alpha^{-2} \sum_{i=1}^{\infty} s_{i,n}^2/\alpha \leq \alpha^{-2} \left(\sum_{i=1}^{\infty} s_{i,n}^2\right)^2 = n^{-2} \eta_n^2.
\]
Together with the bound from the Cauchy–Schwarz inequality for the trace
\[
\text{tr}(H_a \circ H_a^2) \leq \left(\text{tr}(H_a \circ H_a)\right)^{1/2}\left(\text{tr}(H_a^2 \circ H_a^2)\right)^{1/2} = \left(\delta_n^{(1)}\right)^{1/2}\left(\text{tr}(H_a^2 \circ H_a^2)\right)^{1/2},
\]
Lemma 5 gives
\[
\mathbb{P}_n(\delta_n^{(2)} > \epsilon) \leq \mathbb{P}_n(\delta_n^{(1)} > \epsilon \text{ or } \text{tr}(H_a^2 \circ H_a^2) > \epsilon) \leq \frac{3\psi}{\alpha \epsilon \alpha^2} \left(\frac{\eta_n^2}{\alpha \epsilon} + \frac{\eta_n^2}{\alpha^4}\right).
\]


which completes the second step.

**The third step: Tail bound for $\delta_n^{(3)}$.** We begin with diagonalizing $X^\top(I - H_a)^2X$. The equation $(I - H_a)X = X - X(X^\top X + naI_p)^{-1}X^\top X = naX(X^\top X + naI_p)^{-1}$ indicates that
\[
X^\top(I - H_a)^2X = (na)^2(X^\top X + naI_d)^{-1}X^\top X(X^\top X + naI_p)^{-1} \\
= (na)^2(X^\top X + naI_d)^{-1}|X^\top X + naI_p|^{-1} - (na)^2(X^\top X + naI_d)^{-1}(na)(X^\top X + naI_p)^{-1} \\
= (na)^2(X^\top X + na)^{-1} - (na)^3(X^\top X + naI_p)^{-2}.
\]

So, the $i$-th eigenvalue of $X^\top(I - H_a)^2X$ is
\[
\lambda_i := \frac{(na)^2}{s^2_{i,n} + na} - \frac{(na)^3}{\left(s^2_{i,n} + na\right)^2} = n\left\{\frac{\alpha^2}{\left(s^2_{i,n}/n + \alpha\right)^2} - \frac{\alpha^3}{\left(s^2_{i,n}/n + \alpha\right)^2}\right\} = n\frac{\alpha^2 s^2_{i,n}/n}{\left(s^2_{i,n}/n + \alpha\right)^2}
\]
and we get
\[
X^\top(I - H_a)^2X = \mathbf{\Lambda} \mathbf{V}^\top,
\]
where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

Let $\hat{\beta}_0 := (\hat{\beta}_{01}, \hat{\beta}_{02}, \ldots, \hat{\beta}_{0p}) = V^\top \beta_0$. Since $\|\hat{\beta}_0\|_2 \leq \max_{1 \leq i \leq n} \|v_i\|_2 \|\beta_0\|_2 = \|\beta_0\|_2 = \rho^{-1/2}$, we have
\[
\hat{\beta}_0^\top X^\top(I - H_a)^2X \hat{\beta}_0 = \sum_{i=1}^n \lambda_i \hat{\beta}_{0i}^2 \leq \frac{b^2 n}{p} \sum_{i=1}^n \frac{\alpha^2 s^2_{i,n}/n}{\left(s^2_{i,n}/n + \alpha\right)^2} \leq b^2 n^{-1} \sum_{i=1}^n s^2_{i,n} = b^2 \eta.
\]

Then, the Cauchy-Schwarz inequality for the trace gives
\[
\delta_n^{(3)} \leq \left(\text{tr}(H_a \circ H_a)\right)^{1/2} \left(\text{tr}((I - H_a)X \beta_0) \circ \beta_0^2 \circ ((I - H_a)X \beta_0) \circ \beta_0^2\right)^{1/2}/\sigma_0^2 = (\delta_n^{(1)})^{1/2} b^2 \eta/\sigma_0^2.
\]

This, together with (17), yields
\[
\mathbb{P}_n(\delta_n^{(3)} > \epsilon) \leq \mathbb{P}_n(\delta_n^{(1)} > \epsilon \sigma_0^4/\eta^2) \leq \left(\epsilon^2 \frac{\sigma_0^4}{b^4 \eta^2}\right)^{-1} \frac{3 \psi \eta^2}{n \epsilon^2} \leq \frac{3 \psi \eta^2 b^4}{n \epsilon^2 \sigma_0^4},
\]
which completes the third step.

Lastly, combining (17)–(19) proves the assertion
\[
\mathbb{P}_n \left(|\text{FV}(\alpha(a)) - \Delta(a)| > \epsilon\right) \leq \mathbb{P}_n(\delta_n^{(1)} > \epsilon/3) + \mathbb{P}_n(\delta_n^{(2)} > \epsilon/3) + \mathbb{P}_n(\delta_n^{(3)} > \epsilon/3)
\]
\[
\leq \frac{9\psi \eta^2}{n \epsilon} \left(\frac{\eta^4}{\alpha^2} + \frac{\psi^4}{\alpha^4}\right) + 27 \psi \eta^4 b^4 \frac{1}{n \epsilon^2 \sigma_0^4}
\]
\[
\leq 27 \left(\frac{1}{n \epsilon} \psi \eta^2 + \frac{1}{n \epsilon} \psi \eta^4 + \frac{1}{n \epsilon} \psi \eta^4 b^4\right),
\]
which completes the proof.

### 6 Conclusion

In this paper, we considered a Gibbs generalization gap estimation for overparameterized models. We proved that the functional variance (FV) is asymptotically unbiased estimator, even in an overparameterization setting. We proposed a Langevin approximation of FV (LFV) for efficient computation and applied LFV to overparameterized linear regression and non-linear neural network models.

### Acknowledgement

A. Okuno is supported by JST CREST (JPMJCR21N3) and JSPS KAKENHI (21K17718). K. Yano is supported by JST CREST (JPMJCR1763) and JSPS KAKENHI (19K20222, 21H05205).
References

Adlam, B. and Pennington, J. (2020). The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In Proceedings of the 37th International Conference on Machine Learning, pages 74–84.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19:716–723.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. Neural computation, 10:251–276.

Amari, S.-I. and Douglas, S. C. (1998). Why natural gradient? In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 1213–1216.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. (2019). On exact computation with an infinitely wide neural net. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, page 8141?8150.

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

Baum, E. and Haussler, D. (1989). What size net gives valid generalization? Neural computation, 1:151?160.

Belkin, M., Hsu, D., and Xu, J. (2020). Two models of double descent for weak features. SIAM Journal on Mathematics of Data Science, 2:1167–1180.

Belkin, D. Hsu, P. M. (2018). Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, page 2306?2317.

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:1805.01648.

Chiyuan, Z., Samy, B., Moritz, H., Benjamin, R., and Oriol, V. (2017). Understanding deep learning requires rethinking generalization. In Proceedings of the 5th International Conference on Learning Representations.

Dalalyan, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. Journal of the Royal Statistical Society: Series B, 79:651–676.

Devroye, L. and Wagner, T. (1979). Distribution-free inequalities for the deleted and holdout error estimates. IEEE Transactions of Information Theory, 25:202–207.

d’Ascoli, S., Sagun, L., and Biroli, G. (2020). Triple descent and the two kinds of overfitting: where & why do they appear? In Proceedings of the 34th International Conference on Neural Information Processing Systems, pages 3058–3069.

Eaton, M. L. (1989). Group invariance applications in statistics. Institute of Mathematical Statistics.

Gao, T. and Jojic, V. (2016). Degrees of freedom in deep neural networks. In Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence, pages 232–241.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. arXiv:1903.08560.

Jacot, A., Gabriel, E., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 8580–8589.
Karakida, R., Akaho, S., and Amari, S.-i. (2019). Universal statistics of fisher information in deep neural networks: Mean field approach. In *Proceedings of the 32nd International Conference on Artificial Intelligence and Statistics*, pages 1032–1041.

Kearns, M. and Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, 11:1427–1453.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

Mallows, C. (1973). Some comments on $C_p$. *Technometrics*, 15:661–675.

Mandt, S., Hoffman, M. D., and Blei, D. M. (2017). Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18:1–35.

Mario, G., Arthur, J., Stefano, S., Franck, G., Levent, S., Stepha, d., Giulio, B., Clement, H., and Matthieu, W. (2020). Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, page 023401.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63:117–126.

Shibata, R. (1989). *Statistical aspects of model selection*, pages 215–240. Springer-Verlag, New York.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151.

Moody, J. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, pages 847–854.

Sato, I. and Nakagawa, H. (2014). Approximation analysis of stochastic gradient langevin dynamics by using Fokker-Planck equation and Ito process. In *Proceedings of the 31st International Conference on Machine Learning*, pages 982–990.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151.
Takeuchi, K. (1976). Distribution of information statistics and validity criteria of models. *Mathematical Science*, 153:12–18.

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.

Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 17:264–280.

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

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131.
Supplementary material:
A generalization gap estimation for overparameterized models via Langevin functional variance
Akifumi Okuno and Keisuke Yano

A Proof of Lemma 2
We first provide useful formulae of the multivariate normal distribution and then prove Lemma 2.

A.1 Moments of Multivariate Normal Distribution
Let \( z \) be a random vector following a \( p \)-variate normal distribution \( N_p(\mathbf{m}, \Sigma) \), and let \( \tilde{z} := z - \mathbf{m} \sim N_p(\mathbf{0}, \Sigma) \) be a centered random vector. Let \( \mathbf{a} \in \mathbb{R}^p \) be any vector and \( \mathbf{B} \in \mathbb{R}^{p \times p} \), \( \mathbf{C} = (c_{ij}) \in \mathbb{R}^{p \times p} \) be any matrices, for any \( m \in \mathbb{N} \). Then, the following hold:

- (The second order moment): We have
  \[
  E[\|\mathbf{a} - \mathbf{B} \mathbf{z}\|_2^2] = \text{tr}(\mathbf{a} \mathbf{a}^\top) - 2\text{tr}(\mathbf{B} E[\mathbf{z}] \mathbf{a}^\top) + E[\mathbf{z} \mathbf{z}^\top | \mathbf{B}] = \|\mathbf{a} - \mathbf{B} \mathbf{m}\|_2^2 + \text{tr}(\mathbf{B} \Sigma \mathbf{B}^\top),
  \]
  since \( \mathbf{z} \) satisfies \( E[\mathbf{z} \mathbf{z}^\top] = \mathbf{m} \mathbf{m}^\top + \Sigma \).

- (The third order moment): We have, for \( \mathbf{x}^3 := \mathbf{x} \mathbf{x}^\top \mathbf{x} \),
  \[
  E[\langle \mathbf{a}, (\mathbf{B} \mathbf{z})^3 \rangle] = \langle \mathbf{a}, \mathbf{B} E[\mathbf{z} \mathbf{z}^\top \mathbf{B}] \mathbf{z} \rangle = 0,
  \]
  since \( E[\mathbf{z} \mathbf{z}^\top \mathbf{C} \mathbf{z}] = \mathbf{0} \) as its \( i \)-th entry is
  \[
  E \left[ \sum_{j,k=1}^p c_{jk} \tilde{z}_i \tilde{z}_j \tilde{z}_k \right] = \sum_{j,k=1}^p c_{jk} E[\tilde{z}_i \tilde{z}_j \tilde{z}_k] = 0.
  \]

- (The fourth order moment): We have
  \[
  E[\|\mathbf{B} \mathbf{z}\|_2^4] = E[\mathbf{z}^\top \mathbf{B}^\top \mathbf{B} \mathbf{z} \mathbf{B} \mathbf{z}] = 2\text{tr}(\langle \mathbf{B} \mathbf{B} \mathbf{z} \rangle^2) + (\text{tr}(\mathbf{B} \mathbf{B} \Sigma))^2,\]
  where the last equality follows from Section 8.2.4. of Petersen and Pedersen (2012).

A.2 Proof of Lemma 2
The proof consists of two parts: deriving an explicit form of \( \Delta(\alpha) \) and deriving that of \( \nabla_\beta \log f(y_i | x_i, \hat{\beta}) \).

The first part: Explicit form of \( \Delta(\alpha) \). Decompose \( \Delta(\alpha) \) as \( \omega_1(\alpha) - \omega_2(\alpha) \), where \( \omega_1(\alpha) := E_{\beta,y,y^*}(\|y^* - X \hat{\beta}\|_2^2) \) and \( \omega_2(\alpha) := E_{\beta,y}(\|y - X \hat{\beta}\|_2^2) \).

From the definition of \( \omega_1(\alpha) \), we have
\[
\omega_1(\alpha) := E_{\beta,y,y^*}(\|y^* - X \hat{\beta}\|_2^2) = E_{\beta,y}(n \sigma_0^2 + \|X(\hat{\beta}_0 - \beta)\|_2^2) = n \sigma_0^2 + \text{tr}\left\{X^\top X E_{\beta,y}(\{\beta_0 - \hat{\beta}_0\}^2)\right\}.
\]
Since the quasi-posterior of \( \beta \) is \( \mathcal{N}_{\hat{\beta}_0}(\hat{\beta}_0, \boldsymbol{Q}_0) \) with \( \hat{\beta}_0 \) and \( \boldsymbol{Q}_0 \) defined in (2) and (3), respectively, we have
\[
\omega_1(\alpha) = n \sigma_0^2 + n \cdot \text{tr}\{\boldsymbol{Q}_0 + E_y\{\{\beta_0 - \hat{\beta}_0\}^2\}\}\]
\[
= n \sigma_0^2 + n \cdot \text{tr}\{\boldsymbol{Q}_0 + E_y\{\{\beta_0 - \sigma_0^{-2} \boldsymbol{Q}_a X^\top y\}^2\}\},
\]
where \( \boldsymbol{G} = n^{-1} X^\top X \). This, together with (20), gives
\[
\omega_1 = n \sigma_0^2 + n \cdot \text{tr}(\boldsymbol{G} \boldsymbol{Q}_a) + n \cdot \text{tr}(\boldsymbol{G}(\{\beta_0 - \sigma_0^{-2} \boldsymbol{Q}_a X^\top X \beta_0\}^2 + \sigma_0^{-2} \boldsymbol{Q}_a X^\top X \boldsymbol{Q}_a))
\]
Using $G$, we have

$$
\omega_2(a) = n \cdot \text{tr}(GQ_a) + n^2 \sigma_0^2 \text{tr}((GQ_a)^2) = n \cdot \text{tr}(GQ_a) + n^2 \sigma_0^2 \text{tr}((GQ_a)^2) + \text{tr}(XE_a \beta_0)^2).
$$

Further, using $G$, we have

$$
\omega_2(a) = n \cdot \text{tr}(GQ_a) + n^2 \sigma_0^2 \text{tr}((GQ_a)^2) + \text{tr}(XE_a \beta_0)^2) = n \cdot \text{tr}(GQ_a) + n^2 \sigma_0^2 \text{tr}((GQ_a)^2) + \text{tr}(XE_a \beta_0)^2).
$$

Therefore, we get

$$
\Delta(a) = \frac{1}{2 \sigma_0^2} (\omega_1(a) - \omega_2(a)) = \frac{1}{2 \sigma_0^2} \cdot 2n \cdot \text{tr}(GQ_a) = \text{tr}(H_a),
$$

which completes the first part.

The second part: Explicit form of $\nabla_\beta \log f(y_i | x_i, \beta)$. The second part is divided into three subparts (A,B,C).

Subpart (A) of the second part. For $m \in \mathbb{N}$, let

$$
g_{m,i} := E_\beta [x_i ^\top \beta]^m.
$$

Using $g_{m,i}$, we write the posterior expectation of $\log p(x_i, y_i | \beta)$ as

$$
E_\beta [\log f(y_i | x_i, \beta)] = \frac{1}{2 \sigma_0^2} \left[ E_\beta [(x_i ^\top \beta)^2] - 2 y_i E_\beta [x_i ^\top \beta] + y_i^2 \right] + C
$$

where $C$ is a constant independent of $\beta$ and $x_i$. Then, we can write the posterior variance as

$$
\nabla_\beta [\log f(y_i | x_i, \beta)] = \frac{1}{(2 \sigma_0^2)^2} \left[ E_\beta \left[ \{ (y_i - x_i ^\top \beta)^2 - (g_{2,i} - 2y_i g_{1,i} + y_i^2) \}^2 \right] \right]
$$

Expanding the square yields

$$
\nabla_\beta [\log f(y_i | x_i, \beta)] = \frac{1}{(2 \sigma_0^2)^2} \left[ g_{4,i} + (4y_i g_{3,i} + (4y_i^2 - 2g_{2,i} - 2y_i g_{1,i})) g_{2,i} \right].
$$
which completes subpart (A).

**Subpart (B) of the second part.** We next find the expressions for $E_y[h_{2,i}, y_i^2]$, $E_y[h_{1,i}, y_i]$, and $E_y[h_{0,i}]$. From Section 8.2 of Petersen and Pedersen (2012), we have

\[
g_{1,i} = x_i^T \hat{\beta}_a,
\]

\[
g_{2,i} = x_i^T E_x (\hat{\beta} x_i x_i^T \hat{\beta} x_i) = x_i^T (Q_a + \hat{\beta} a^2) x_i,
\]

\[
g_{3,i} = x_i^T E_x (\hat{\beta} x_i x_i^T \hat{\beta} x_i) x_i = x_i^T (Q_a + \hat{\beta} a^2) x_i x_i^T (Q_a + \hat{\beta} a^2) x_i + x_i^T \hat{\beta}_a (Q_a - \hat{\beta} a^2) x_i,
\]

\[
g_{4,i} = 3x_i^T (Q_a + \hat{\beta} a^2) x_i x_i^T (Q_a + \hat{\beta} a^2) x_i - 2(x_i^T \hat{\beta}_a)^4.
\]

These lead to the following expressions for $h_{m,i} (m = 0, 1, 2)$:

\[
h_{2,i} = 4g_{2,i} - 4g_{1,i} = 4[x_i^T Q_a x_i],
\]

\[
h_{1,i} = 4g_{1,i} g_{2,i} + 4g_{3,i} = -8(x_i^T \hat{\beta}_a) (x_i^T Q_a x_i),
\]

\[
h_{0,i} = g_{4,i} - g_{2,i}^2 = 2(x_i^T (Q_a + \hat{\beta} a^2) x_i - 2(x_i^T \hat{\beta}_a)^4 = 4(x_i^T \hat{\beta}_a)^2 [x_i^T Q_a x_i] + 2 [x_i^T Q_a x_i]^2.
\]

Thus, using a vector $e_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0) \in \{0, 1\}^n$ whose $i$-th entry is 1 and 0 otherwise, we get

\[
E_y[h_{2,i}, y_i^2] = 4 [x_i^T Q_a x_i] E_y[y_i^2] = 4 [x_i^T Q_a x_i] [\sigma_0^2 + (x_i^T \hat{\beta}_a)^2],
\]

\[
E_y[h_{1,i}, y_i] = -8 \sigma_0^{-2} x_i^T Q_a x_i E_y[y y^T] e_i (x_i^T Q_a x_i) \quad (\because \hat{\beta}_a = \sigma_0^{-2} Q_a X^T y) \text{ and } y_i = y_i^2 e_i
\]

\[
= -8 \sigma_0^{-2} x_i^T Q_a x_i \sigma_0^2 I_n + (X \hat{\beta}_0 a^2) e_i (x_i^T Q_a x_i)
\]

\[
= -8 [x_i^T Q_a x_i]^2 - 8 \sigma_0^{-2} x_i^T Q_a X \hat{\beta}_0 a x_i x_i^T Q_a x_i \quad (\because x_i^T e_i = x_i),
\]

and

\[
E_y[h_{0,i}] = 4 \sigma_0^{-4} x_i^T Q_a X^T E_y[y y^T] X Q_a x_i \{x_i^T Q_a x_i] + 2 [x_i^T Q_a x_i]^2
\]

\[
= 4 \sigma_0^{-4} x_i^T Q_a X^T [\sigma_0^2 I_n + (X \hat{\beta}_0 a^2) X Q_a x_i] x_i^T Q_a x_i] + 2 [x_i^T Q_a x_i]^2
\]

\[
= 4 \sigma_0^{-4} x_i^T Q_a X^T Q_a x_i x_i^T Q_a x_i x_i^T Q_a x_i + 2 [x_i^T Q_a x_i]^2,
\]

which completes subpart (B).

**Subpart (C) of the second part.** We lastly derive an explicit expression of $\sum_{i=1}^n \log f(y_i | x_i, \beta)$. For any matrices $A = (a_{ij})$ and $B = (b_{ij}) \in \mathbb{R}^{d \times d}$, we denote by $L_1(A)$

\[
L_1(A) := \sum_{i=1}^n x_i^T A x_i = \text{tr}[X A X^T]
\]
and denote by $L_2(A, B)$

$$L_2(A, B) := \sum_{i=1}^{n} x_i^T A x_i x_i^T B x_i = \text{tr}((XAX^\top) \circ (XBX^\top)).$$

From (23) and the identity $XQ_aX^\top = \sigma_0^2 H_a$, we get

$$\sum_{i=1}^{n} E_y[h_{1,i} y_i] = 4\sigma_0^2 L_1(Q_a) + 4L_2(Q_a, \beta_0^2) = 4\sigma_0^2 \text{tr}[H_a] + 4\sigma_0^2 \text{tr}[H_a \circ (X\beta_0)^2].$$

From (24), we get

$$\sum_{i=1}^{n} E_y[h_{0,i}] = -8L_2(Q_a, Q_a) - 8\sigma_0^{-2} L_2(Q_aX^\top X\beta_0\beta_0^\top Q_a) = -8\sigma_0^2 \text{tr}[H_a] - 8\sigma_0^2 \text{tr}[H_a \circ (X\beta_0)^2].$$

Likewise, we have

$$\sum_{i=1}^{n} E_y[h_{0,i} | y_i] = 4\sigma_0^{-2} L_2(Q_aX^\top XQ_a, Q_a) + 4\sigma_0^{-4} L_2(Q_aX^\top X\beta_0\beta_0^\top X^\top XQ_a, Q_a) + 2L_2(Q_a, Q_a) = 4\sigma_0^4 \text{tr}[H_a^2] + 4\sigma_0^2 \text{tr}[H_a(X\beta_0)^2] + 2\sigma_0^2 \text{tr}[H_a \circ H_a].$$

These identities yield

$$E_y[\mathbb{F}V(a)] = E_y \left[ \sum_{i=1}^{n} \nabla_{\beta} \log f(y_i | x_i, \beta) \right]$$

$$= E_y \left[ \sum_{i=1}^{n} \frac{1}{2(\sigma_0^2)} [h_{1,i} y_i^2 + h_{1,i} y_i + h_{0,i}] \right]$$

$$= \frac{1}{2(\sigma_0^2)} \left[ \sum_{i=1}^{n} E_y[h_{1,i} y_i^2] + \frac{n}{n} E_y[h_{1,i} y_i] + \frac{n}{n} E_y[h_{0,i}] \right]$$

$$= \text{tr}[H_a] - \frac{3}{2} \text{tr}[H_a \circ H_a] + \text{tr}[H_a \circ H_a^2]$$

$$+ \frac{1}{\sigma_0^2} \text{tr}[H_a \circ (X\beta_0)^2 - 2\sigma_0^2 H_a \circ (X\beta_0)^2 + (H_a \circ (X\beta_0)^2)] = (I_a - H_a)(X\beta_0)^2.$$

which completes subpart (C) of the second part and concludes the proof. \hfill \Box

## B Proof of Proposition 1

Since we have, for some constant $C$ independent of $\beta$,

$$E_{\beta}[\log f(y | X, \beta)] = -\frac{1}{2\sigma_0^2} \|y - X\hat{\beta}_a\|^2_2 - \frac{1}{2\sigma_0^2} \text{tr}[XQ_aX^\top] + C,$$

we get

$$\nabla_{\beta}[\log f(y | X, \beta)] = \frac{1}{2(\sigma_0^2)} E_{\beta} \left[ \|y - X\beta\|^2_2 - \|y - X\hat{\beta}_a\|^2_2 - \text{tr}[XQ_aX^\top] \right]^2$$

$$= \frac{1}{2(\sigma_0^2)} E_{\beta} \left[ \|X\hat{\beta}\|^2_2 - \langle m, X\hat{\beta} \rangle - \zeta \right]^2,$$

where $\hat{\beta} := \beta - \hat{\beta}_a, m := 2(y - X\hat{\beta}_a), \zeta := \text{tr}(XQ_aX^\top)$. Expanding the square gives

$$\nabla_{\beta}[\log f(y | X, \beta)] = \frac{1}{2(\sigma_0^2)} E_{\beta} \left[ \|X\hat{\beta}\|^2_2 - 2\langle m, X\hat{\beta} \rangle + \hat{\beta}_a + \hat{\beta}_a^\top (mm^\top - 2\zeta I) X\hat{\beta} + 2\zeta \langle m, X\hat{\beta} \rangle + \zeta^2 \right].$$
As \( \hat{\beta} \) is centered, using (21) and (22) leads to

\[
\forall \beta \left[ \log f(y | X, \beta) \right] = \frac{1}{(2\sigma_0^2)^2} \mathbb{E}_\beta [\|X\hat{\beta}\|^2 + \text{tr}(X^T (mm^T - 2\zeta I)X)\hat{\beta}\hat{\beta}^T + \zeta^2] \\
= \frac{1}{(2\sigma_0^2)^2} \left\{ 2\text{tr}(XQ_aX^T)^2 + (\text{tr}(XQ_aX^T))^2 + \text{tr}((mm^T - 2\zeta I)XQ_aX^T + \zeta^2) \right\}.
\]

From equations \( m = 2(I - H_a)y, XQ_aX^T = \sigma_0^2 H_a \) and \( \zeta = \text{tr}(XQ_aX^T) \), we have

\[
\mathbb{E}_y [\forall \beta \left[ \log f(y | X, \beta) \right]] = \frac{1}{(2\sigma_0^2)^2} \left\{ 2\text{tr}(\sigma_0^2 H_a)^2 + \zeta^2 + \sigma^2 \text{tr}(\mathbb{E}_y [mm^T] H_a) - 2\zeta^2 + \zeta^2 \right\} \\
= \frac{1}{\sigma_0^2} \text{tr}(H_a^2) + \frac{1}{\sigma_0^2} \text{tr}(\mathbb{E}_y [(I - H_a)y](X\hat{\beta}_0)^\sigma_2) + (\zeta^2 - 2\zeta^2 + \zeta^2) \\
= \frac{1}{\sigma_0^2} \text{tr}(H_a^2) + \frac{1}{\sigma_0^2} \text{tr}((I - H_a)H_a(I - H_a)H_a) \\
= \text{tr}(H_a^2)/2 + (I - H_a)H_a + \frac{1}{\sigma_0^2} \text{tr}((I - H_a)H_a(I - H_a)(X\hat{\beta}_0)^\sigma_2) \\
= \text{tr}(H_a)^3/2 + \text{tr}(H_a)^3 + \frac{1}{\sigma_0^2} \text{tr}(H_a((I - H_a)(X\hat{\beta}_0)^\sigma_2),
\]

which proves the assertion. 

\( \square \)

C Supporting Propositions

**Proposition 2** (Cheng et al. (2018) Theorem 2). Let \( R > 0 \) and let let \( 0 < \varepsilon \leq \frac{p}{\sqrt{x^2 + \lambda_{\text{max}}(Q_a^{-1}) + 1}} \) be a desired accuracy. Assume that \( \exp(-R^2\lambda_{\text{max}}(Q_a^{-1})) \leq 2R^2 \lambda_{\text{min}}(Q_a^{-1}) \) and \( \|Y^{(0)} - \hat{\beta}\|_2 \leq R \). Then, it holds for \( \delta := \frac{\varepsilon}{2^\rho \lambda_{\text{max}}(Q_a^{-1})} \) and \( t = \exp(\frac{3R^2 \lambda_{\text{max}}(Q_a^{-1})}{\varepsilon^2}) \) that

\[
W_1(\mathcal{D}(\hat{Y}_{\lambda}), \mathcal{D}(Y^{(0)})) \leq \varepsilon,
\]

where \( W_1(\cdot, \cdot) \) denotes 1-Wasserstein distance (a.k.a. Earth-mover’s distance).

**Proof.** Substituting \( m := 2\lambda_{\text{min}}(Q_a^{-1}) \) and \( L := 2\lambda_{\text{max}}(Q_a^{-1}) \) to Cheng et al. (2018) Theorem 2 proves the assertion. 

\( \square \)