Optimizing Information-theoretical Generalization Bound via Anisotropic Noise in SGLD

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, the information-theoretical framework has been proven to be able to obtain non-vacuous generalization bounds for large models trained by Stochastic Gradient Langevin Dynamics (SGLD) with isotropic noise. In this paper, we optimize the information-theoretical generalization bound by manipulating the noise structure in SGLD. We first prove that with constraint to guarantee low empirical risk and commonly used data-dependent priors, the optimal noise covariance of SGLD in terms of the bound is similar to the empirical gradient covariance. We further prove that if the generalization bound is jointly optimized with respect to both prior and posterior, the optimal noise covariance is the square root of the expected gradient covariance. Technically, we develop new information-theoretical bounds that enable such an optimization analysis. We then apply matrix analysis to derive the form of optimal noise covariance. Presented constraint and results are validated by the empirical observations.

1 Introduction

Generalization ability is one of the core questions in learning theory [4], but remains unclear for deep learning models [16, 35]. Existing generalization bounds based on the capacity control become vacuous for practical deep learning models due to the over-parameterization property [1, 25, 36].

From information theoretical perspective, recent work [29, 34] bound the generalization error by the mutual information between the dataset and the learned parameters. This result reveals that good generalization occurs when the learned parameters do not depend on a specific dataset too much, which is intuitively reasonable and closely related with the idea of algorithm stability [6, 14, 18, 23, 19] and differential privacy [8, 10]. Meanwhile, based on information-theoretic metric, it allows to analyze general classes of updates and models, e.g., stochastic iterative algorithms for non-convex objectives, and hence is applicable to deep learning. It has been shown that the information theoretical bounds can achieve non-vacuous bounds for large/deep models [9, 36].

Notably, the information theoretical bound is realized in [27] via decomposing the mutual information across iterations. This framework can perform a step-wise analysis of Stochastic Gradient Langevin Dynamics (SGLD) [31, 28], by evaluating the mutual information conditional on the previous learned parameters at each step. The bound is further improved in [24, 13] by bounding the mutual information with the KL divergence between prior and posterior distribution, and carefully choosing a data-dependent prior. Specifically, at each iteration, it uses part of the data to estimate a prior of the gradient and demonstrates that the bound captures the generalization behavior of neural network.

In this paper, we go beyond the results in [24, 13] to further optimize the information theoretical generalization bound. Specifically, we propose to optimize the structure of the noise in SGLD such that the generalization bound is minimized while a low empirical risk is guaranteed. To this end, we

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

first show that the trace of the noise covariance in SGLD is a valid constraint that governs the empirical risk behavior both theoretically and empirically. We further argue that the generalization bounds [24, 13] do not allow us to analyze the case of state-dependent noise due to their multi-dependence of each step's noise covariance. Then we devise two new information theoretical generalization bounds that are parallel to those bounds in [24], but facilitate the analysis under two different prior choices.

With these technical preparation, we first prove that with the constraint on the trace of the noise covaraince and the commonly used data-dependent priors, the optimal noise covariance of SGLD exhibits structure similar to the empirical gradient covariance. We further prove that when jointly optimizing both the prior and the posterior, the optimal noise covariance is the square root of the expected gradient covariance and the optimal prior recovers the prior in [24].

Our result on the optimal noise structure lends support to the belief that the noise introduced by SGD is superior to the isotropic noise, which has been widely observed [17, 33, 37, 35]. Previous work [22, 37] have provided justification for this effect from the perspective of the escaping rate. Our result verifies the benefit of SGD noise in terms of optimizing the generalization bound.

Specifically, our contribution can be summarized as follows:

1. We formulate two problems to analyze the effect of noise covariance through information-theoretical framework, and further demonstrate the difficulty to analyze the effect using existing information-theoretical generalization bounds;
2. We then propose two information-theoretical generalization bounds which are parallel to existing bounds, but facilitate the analysis of the two problems respectively;
3. We obtain the optimal structures of noise covariances for these two problems, and demonstrate their similarity to empirical/expected gradient covariance;
4. We conduct empirical experiments to furhter support our theoretical results.

1.1 Related Works

Information-theoretical bounds. Recently, researchers [34, 29, 27] propose to bound the generalization error by the mutual information between output hypothesis and input samples. Negrea et al. [24] further tighten the bound by designing a data-dependent prior. Following [24], Haghifam et al. [13] obtain comparable results through conditional mutual information [30]. Other related work [2, 15, 5, 12] tightens the information theoretical generalization bounds from different perspectives. There are also high probability generalization bound [23, 25, 20] obtained by combining information theory [7] and PAC Bayesian framework [26, 31].

Effects of noise covariance. One recent work [37] demonstrates empirically that even with well-tuned scaling, isotropic SGLD still has much worse generalization error than mini-batch SGD. They further obtain the optimal covariance matrix for best escaping efficiency at local minima. Furthermore, [32] show empirically large-batch SGD with diagonal Fisher Gaussian noise can recover similar validation performance as small-batch SGD. It has also been proven [22] that the stationary distribution of state-dependent SGLD with the empirical loss approximated by quadratic function obeys the power law, and has a better escaping efficiency than the isotropic noise SGLD.

2 Preliminaries

Notations. For a positive integer N, we use $[N]$ to denote the index set $\{1, 2, \cdots, N\}$, and use $\boldsymbol{V}_{[T]}$ to denote $\{\boldsymbol{V}_t\}_{t=1}^T$ where $\{\boldsymbol{V}_t\}_{t=1}^T$ is a series of elements. For a finite set $\boldsymbol{S} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N\}$, $|\boldsymbol{S}|$ is the cardinality of $\boldsymbol{S}$, and $\boldsymbol{S}_J \triangleq \{\boldsymbol{z}_i\}_{i \in J}$ is a subset of $\boldsymbol{S}$ with indices in set $J \subset [N]$. Also, $\boldsymbol{z} \sim \boldsymbol{S}$ denotes that $\boldsymbol{z}$ is uniformly sampled from the set $\boldsymbol{S}$. We use $\mathbb{I}_{d\times d}$ for the $d \times d$ identity matrix, abbreviated as $\mathbb{I}$ when dimension is clear. For a differentiable function f, we denote the gradient of f at point $\boldsymbol{W}$ as $\nabla f(\boldsymbol{W})$. For two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, we denote the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$ as $\mathbb{P}^{\boldsymbol{Y}}(\boldsymbol{X})$ and denote the conditional expectation of $\boldsymbol{X}$ given $\boldsymbol{Y}$ as $\mathbb{E}^{\boldsymbol{Y}}(\boldsymbol{X})$. For a matrix M, $M \succeq 0$ means that M is positive semi-definite.

Supervised Learning. In this paper, we focus on the supervised learning. It conducts the empirical risk minimization (ERM) over training data: $\mathrm{Minimize}_{\boldsymbol{W}}\, \mathcal{R}_{\boldsymbol{S}}(\boldsymbol{W}) \triangleq \frac{1}{|\boldsymbol{S}|}\sum_{\boldsymbol{z}\in\boldsymbol{S}} \ell(\boldsymbol{z}; \boldsymbol{W})$, where $\boldsymbol{W} \in \mathbb{R}^d$ is the parameter of the model, $\boldsymbol{S}$ is the training set with each data point i.i.d. sampled from

a distribution $\mathcal{D}$, ℓ is the (individual) loss function. A *stochastic algorithm* solves the ERM by outputting a distribution $P^{\boldsymbol{S}}$ of parameter $\boldsymbol{W}$. The *generalization error* measures the gap between the population risk and the training risk as

$$\mathrm{Gerr} \triangleq \mathbb{E}_{\boldsymbol{S},P^{\boldsymbol{S}}}\left[\mathcal{R}_{\mathcal{D}}(\boldsymbol{W}) - \mathcal{R}_{\boldsymbol{S}}(\boldsymbol{W})\right]. \tag{1}$$

SGD and SGLD. A commonly used stochastic algorithm is Stochastic Gradient Descent (SGD), whose update rule at step t ($t = 1, 2, 3, \cdots$) is defined as

$$\boldsymbol{W}_t = \boldsymbol{W}_{t-1} - \eta_t \nabla \mathcal{R}_{\boldsymbol{S}_{\boldsymbol{V}_t}}(\boldsymbol{W}_{t-1}), \tag{2}$$

where $\boldsymbol{V}_t$ is sampled uniformly without replacement from $[N]$ with size b_t. Given $\boldsymbol{W}_{t-1}$, $\boldsymbol{W}_t$ is random with

$$\mathbb{E}\left[\boldsymbol{W}_t|\boldsymbol{W}_{t-1}\right] = \boldsymbol{W}_{t-1} - \eta_t \nabla \mathcal{R}_{\boldsymbol{S}}\left(\boldsymbol{W}_{t-1}\right), \ \mathrm{Cov}\left[\boldsymbol{W}_t|\boldsymbol{W}_{t-1}\right] = \frac{N-b_t}{b_t(N-1)}\boldsymbol{\Sigma}^{sd}_{\boldsymbol{S},\boldsymbol{W}_{t-1}},$$

where $\boldsymbol{\Sigma}^{sd}_{\boldsymbol{S},\boldsymbol{W}_{t-1}} \triangleq \mathrm{Cov}_{\boldsymbol{z}\sim\boldsymbol{S}}\left[\nabla \mathcal{R}_{\boldsymbol{z}}(\boldsymbol{W}_{t-1})\right]$ is the covariance of $\nabla \mathcal{R}_{\boldsymbol{z}}$ with $\boldsymbol{z}$ single drawn from $\boldsymbol{S}$. The superscript ^{sd} means "single draw". Similarly, we define the population covariance of gradient as $\boldsymbol{\Sigma}^{\mathrm{pop}}_{\boldsymbol{W}} \triangleq \mathrm{Cov}_{\boldsymbol{z}\sim\mathcal{D}}\left[\nabla \mathcal{R}_{\boldsymbol{z}}(\boldsymbol{W})\right]$. We define the Hessian of $\mathcal{R}_{\boldsymbol{S}}(\boldsymbol{W})$ at point $\boldsymbol{W}$ as $\mathcal{H}_{\boldsymbol{S},\boldsymbol{W}}$.

The Stochastic Gradient Langevine Dynamics (SGLD) is given by

$$\boldsymbol{W}_t = \boldsymbol{W}_{t-1} - \eta_t \nabla \mathcal{R}_{\boldsymbol{S}_{\boldsymbol{V}_t}}(\boldsymbol{W}_{t-1}) + \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1})), \tag{3}$$

where $\boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1}) \in \mathbb{R}^{d\times d}$ is a positive semi-definite matrix with dependence on $\boldsymbol{S}$ and $\boldsymbol{W}_{t-1}$. SGLD (Eq. (38) with $\boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1}) = c_t\mathbb{I}$ where c_t a positive constant, is called *isotropic* SGLD, and SGLD with $\boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1})$ dependent on $\boldsymbol{W}_{t-1}$ is called *state-dependent* SGLD.

Statistics for iterative algorithms. Both SGD and SGLD are *stochastic iterative algorithms* [11]. Specifically, the update rule at step t of a stochastic iterative algorithm $\mathcal{A}$ can be generally characterized as $\boldsymbol{W}_t = \mathcal{M}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}, \boldsymbol{V}_t)$, where $\boldsymbol{V}_t$ is the *auxiliary random variable* at step t, e.g., $\boldsymbol{V}_t$ in Eq.(2) and Eq.(3). We use $P^{\boldsymbol{S},\boldsymbol{V}_{[T]}}$ to denote the joint distribution of $(\boldsymbol{W}_t)_{t=0}^T$ conditional on $\{\boldsymbol{S}, \boldsymbol{V}_{[T]}\}$, $P^{\boldsymbol{S},\boldsymbol{V}_{[T]}}_{i:j}$ to denote the joint distribution of $(\boldsymbol{W}_t)_{t=i}^j$ conditional on $\{\boldsymbol{S}, \boldsymbol{V}_{[T]}\}$, and $P^{\boldsymbol{S},\boldsymbol{V}_{[T]}}_{j|j-1}$ to denote the distribution of $\boldsymbol{W}_j$ conditional on $\{\boldsymbol{W}_{j-1}, \boldsymbol{S}, \boldsymbol{V}_{[T]}\}$.

Decomposition of KL divergence. The following Lemma is extensively used throughout this paper.

Lemma 1 (Proposition 2.6, [24]). *Let $Q_{0:T}$ and $P_{0:T}$ are two probability measures on $\mathbb{R}^{d\times(T+1)}$ with $Q_0 = P_0$. Then, the KL divergence between $Q_{0:T}$ and $P_{0:T}$ can be decomposed into*

$$\mathrm{KL}(Q_{0:T}||P_{0:T}) = \sum_{t=1}^{T} \mathbb{E}_{Q_{0:t-1}}\left[\mathrm{KL}\left(Q_{t|t-1}\|P_{t|t-1}\right)\right].$$

Information theoretical generalization bound. Several existing information-theoretical bounds [13, 24] share similar framework. We state one representative proposed in [24].

Proposition 1 (Corollary of Theorem 2.5 in [24]). *Let $\boldsymbol{S}$ be a data set sampled from $\mathcal{D}^N$, and let the loss function $\ell(\boldsymbol{z}; \boldsymbol{W})$ be $[a_1, a_2]$ bounded. Let $\mathcal{A}$ be a stochastic iterative optimization algorithm with update rule $\boldsymbol{W}_t = \mathcal{M}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}, \boldsymbol{V}_t)$, where the start point $\boldsymbol{W}_0$ is initialized according to distribution $\mathcal{W}_0$. Let $\mathcal{B}$ be another stochastic iterative optimization algorithm with update rule $\boldsymbol{W}_t = \tilde{\mathcal{M}}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}_{\boldsymbol{J}}, \boldsymbol{V}_t, \boldsymbol{J})$ and the same initialization distribution $\mathcal{W}_0$, where $\boldsymbol{J}$ is sampled uniformly without replacement from $[N]$ with size $N-1$. Given $\boldsymbol{S}$, $\boldsymbol{J}$, and $\boldsymbol{V}_{[T]}$, denote the joint distribution of $(\boldsymbol{W}_t)_{t=1}^T$ of $\mathcal{A}$ as $Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}$, and the joint distribution of $(\boldsymbol{W}_t)_{t=1}^T$ of $\mathcal{B}$ as $P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}$. Then, the generalization error of $\mathcal{A}$ can be bounded as:*

$$\mathbb{E}_{\boldsymbol{S},\boldsymbol{V}_{[T]}}\left[\mathcal{R}_{\mathcal{D}}\left(Q_T^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\right) - \hat{\mathcal{R}}_S\left(Q_T^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\right)\right] \leq \mathbb{E}_{\boldsymbol{S},\boldsymbol{J},\boldsymbol{V}_{[T]}}\sqrt{\frac{(a_2-a_1)^2}{2}\mathrm{KL}\left(Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\,\|P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}\right)}. \tag{4}$$

This is used to obtain the generalization error bound for SGLD [24, Theorem 3.1] by further combing Lemma 1 and the update rule of SGLD. In Section 4.1, we will argue the difficulty of applying Proposition 1 to analyze the effect of noise structures.

3 Constraint on the Covariance to Guarantee the Empirical Performance

In this section, we derive the constraint of noise covariance $\boldsymbol{\Sigma}_t$ in Eq. (3), under which we will optimize the generalization bounds in the rest of this paper.

Without any constraint on the noise covariance, optimizing generalization error is trivial but meaningless: a direct combination of Theorem 1 of [34] and Theorem 1 of [27] shows that for isotropic SGLD with $\boldsymbol{\Sigma}_t = \sigma_t \mathbb{I}$, the generalization error after T step satisfies $\mathrm{Gerr} \leq \mathcal{O}((\sum_{t=1}^{T} \log(1 + 1/\sigma_t))^{1/2})$. Therefore, as $\sigma_t \to \infty$ for $t \in [T]$, we have $\mathrm{Gerr} \to 0$, but then the update of SGLD is dominated by the noise, leading to arbitrary low convergence rate. Hence, we need constraints on the covariance in order to control the empirical risk when minimizing the generalization error. Specifically, the expected decrease of loss for one iteration can be bounded as follows.

Lemma 2. *Let empirical risk $\mathcal{R}_{\boldsymbol{S}}(\boldsymbol{W})$ be β-smooth w.r.t. $\boldsymbol{W}$. Let $\boldsymbol{W}_{[T]}$ be given by state-dependent SGLD (Eq. (3)). Then,*

$$\mathbb{E}_{t+1|t}\mathcal{R}_{\boldsymbol{S}}(\boldsymbol{W}_{t+1}) \leq \mathcal{R}_{\boldsymbol{S}}(\boldsymbol{W}_t) - \left(1 - \frac{\beta\eta_{t+1}}{2}\right)\eta_{t+1}\left\|\nabla\mathcal{R}_{\boldsymbol{S}}\left(\boldsymbol{W}_t\right)\right\|^2 + \frac{\beta}{2}\operatorname{tr}\left(\frac{\eta_{t+1}^2(N-b_t)}{(N-1)b_t}\boldsymbol{\Sigma}_{\boldsymbol{S},\boldsymbol{W}_t}^{sd} + \boldsymbol{\Sigma}_{t+1}(W_t, \boldsymbol{S})\right).$$

The proof can be obtained by a standard analysis in optimization, and we defer it to Appendix B. By Lemma 2, the noise covariance $\boldsymbol{\Sigma}_t$ affects the upper bound of the empirical risk by its *trace*. Therefore, it is reasonable to keep $\operatorname{tr}(\boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1}))$ unchanged while seeking the optimal $\boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1})$ to minimize the generalization error Gerr. The constraint is given formally as follows:

Constraint 1. *The trace of $\boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1})$ is fixed when optimizing the generalization error. That is, there exists a positive constant $c_t(\boldsymbol{S}, \boldsymbol{W}_{t-1})$ dependent on $\boldsymbol{S}$ and $\boldsymbol{W}_{t-1}$, such that,*

$$\operatorname{tr}(\boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1})) = c_t(\boldsymbol{S}, \boldsymbol{W}_{t-1}).$$

Specifically, with Constraint 1, we optimize the generalization error Gerr within the domain $\{(\boldsymbol{\Sigma}_t)_{t=1}^T : \boldsymbol{\Sigma}_t \succeq 0 \text{ and } \operatorname{tr}(\boldsymbol{\Sigma}_t) = c_t\}$. Similar constraint is also proposed by [37] from the standpoint of kinetic energy when analyzing the effect of noise structure on the escaping efficiency from saddle points.

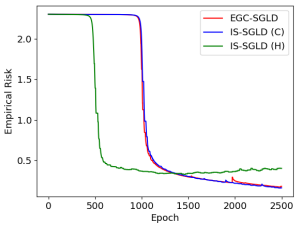


Figure 1: Training error of SGLD with different noise covariances. The experiment is run on the Fashion-MNIST dataset with a four-layer neural network (see Appendix F).

We next verify this constraint empirically. We run SGLD with different covariances on a four-layer neural network for the Fashion-MNIST classification problem. Concretely, the noise covariances are chosen respectively as $\boldsymbol{\Sigma}_t^{(1)} = \boldsymbol{\Sigma}_{\boldsymbol{S},\boldsymbol{W}_{t-1}}^{sd}$ ("EGC-SGLD" in Fig.1) and $\boldsymbol{\Sigma}_t^{(2)} = \frac{1}{d}\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{S},\boldsymbol{W}_{t-1}}^{sd})\mathbb{I}$ ("IS-SGLD (C)" in Fig.1). It is easy to verify that $\operatorname{tr}(\boldsymbol{\Sigma}_t^{(1)}) = \operatorname{tr}(\boldsymbol{\Sigma}_t^{(2)})$, which is exactly Constraint 1. We can see from Fig.1 that the convergence curves corresponding to $\boldsymbol{\Sigma}_t^{(1)}$ and $\boldsymbol{\Sigma}_t^{(2)}$ almost coincide with each other, validating the Constraint 1.

On the other hand, another quantity $\operatorname{tr}(\mathcal{H}_{\boldsymbol{S},\boldsymbol{W}}\boldsymbol{\Sigma}_t)$ has also been proposed to govern the convergence behavior [32, Theorem 4.1]. We plot the convergence curve of SGLD with noise covariance $\boldsymbol{\Sigma}_t^{(3)} = \frac{\operatorname{tr}(\mathcal{H}_{\boldsymbol{S},\boldsymbol{W}_{t-1}}\boldsymbol{\Sigma}_{\boldsymbol{S},\boldsymbol{W}_{t-1}}^{sd})}{\operatorname{tr}(\mathcal{H}_{\boldsymbol{S},\boldsymbol{W}_{t-1}})}\mathbb{I}$ ("IS-SGLD (H)" in Fig.1). While $\operatorname{tr}(\mathcal{H}_{\boldsymbol{S},\boldsymbol{W}}\boldsymbol{\Sigma}_t^{(1)}) = \operatorname{tr}(\mathcal{H}_{\boldsymbol{S},\boldsymbol{W}}\boldsymbol{\Sigma}_t^{(3)})$, there is a significant gap between curves of "EGC-SGLD " and "IS-SGLD (H)" in Fig.1. This implies that $\operatorname{tr}(\mathcal{H}_{\boldsymbol{S},\boldsymbol{W}}\boldsymbol{\Sigma}_t)$ is not a good constraint for the empirical risk of SGLD, supporting the validity of Constraint 1 from another side.

Hence in the rest of this paper, we study the optimal structure of $\boldsymbol{\Sigma}_t$ under the Constraint 1. Naturally, there are two interesting problems.

Problem 1. For the update rule $\tilde{\mathcal{M}}_{[T]}$ fixed as some commonly used prior in existing literature, what is the optimal structure of $\boldsymbol{\Sigma}_t$ under the Constraint 1?

Problem 2. What is the optimal structure of $\boldsymbol{\Sigma}_{[T]}$ under the Constraint 1 if the covariance $\boldsymbol{\Sigma}_{[T]}$ and $\tilde{\mathcal{M}}_{[T]}$ are optimized jointly? (Here, optimizing $\tilde{\mathcal{M}}_{[T]}$ means optimizing the prior distribution of $\boldsymbol{W}_{[T]}$, whose update rules are not necessarily Gaussian mechanisms as those in [24, 13].)

In Section 4, we propose new information-theoretical bounds that are more suitable for the problem discussion than existing Proposition 1. Then Section 5 and Section 6 solve Problem 1 and Problem 2, respectively.

4 New Information-theoretical Bounds

In this section, we establish two new information-theoretical generalization bounds that are parallel to Theorems in [24] but more suitable for analyzing the effect of noise structure. The basic idea of these two bounds is to reverse the order of prior and posterior in the KL divergence to let $\mathbf{\Sigma}_t$ only affect one term in the generalization bound for each t when finding the optimal point.

4.1 Difficulties When Applying Traditional Information-theoretical Bounds

We first show the difficulties of analyzing optimal noise structure based on the existing generalization bound in [24]. By Lemma 1, the generalization error bound in Proposition 1 can be rewritten as

$$\mathbb{E}_{\boldsymbol{S},\boldsymbol{J},\boldsymbol{V}_{[T]}}\sqrt{\frac{(a_2-a_1)^2}{2}\sum_{s=1}^{T}\mathbb{E}_{Q_{s-1}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}}\,\mathrm{KL}\left(Q_{s|(s-1)}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\,\middle\|\,P_{s|(s-1)}^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}\right)}. \tag{5}$$

For **Problem 1**, for each $\mathbf{\Sigma}_i$ where $i \in [T-1]$, the Eq. (5) depends on $\mathbf{\Sigma}_i$ via two terms $\mathbb{E}_{Q_{i-1}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}}\mathrm{KL}(Q_{i|(i-1)}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\|P_{i|(i-1)}^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}})$ and $\sum_{s=i+1}^{T}\mathbb{E}_{Q_{s-1}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}}\mathrm{KL}(Q_{s|(s-1)}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\|P_{s|(s-1)}^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}})$. While the first term is easy to deal with as $Q_{i-1}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}$ is irrelevant of $\mathbf{\Sigma}_i$, the second term depends on $\mathbf{\Sigma}_i$ via the distribution $Q_{s-1}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}$ for $s \geq i+1$, and can be very complex (please refer to Appendix C.1).

For **Problem 2**, there are two difficulties. For a fixed $Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}$, in order to obtain the optimal $P_{s|(s-1)}^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}$, one would first calculate the outside expectation of Eq. (5) over $\boldsymbol{S}_{\boldsymbol{J}^c}$, because $P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}$ is independent of $\boldsymbol{S}_{\boldsymbol{J}^c}$. For each term $\mathbb{E}_{Q_{s-1}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}}\mathrm{KL}\,(Q_{s|(s-1)}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\|P_{s|(s-1)}^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}})$ where $s \in [T]$, both the probability measure $Q_{s-1}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}$ and the function $\mathrm{KL}(Q_{s|(s-1)}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\|P_{s|(s-1)}^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}})$ has dependency on $\boldsymbol{S}_{\boldsymbol{J}^c}$, which makes evaluating the generalization bound with respect to $P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}$ extremely hard. Furthermore, when we come to optimize the bound w.r.t. $\mathbf{\Sigma}_{[T]}$, the same difficulty encountered in **Problem 1** still exists.

4.2 Information-theoretical Generalization Bound for Identifying the Noise Effect

We first present an information-theoretical generalization bound together with an additional assumption to solve **Problem 1**.

Theorem 1. *Let $\boldsymbol{S}$, $\ell(\boldsymbol{z};\boldsymbol{W})$, $\mathcal{A}$, $\mathcal{W}_0$ and $\boldsymbol{V}_{[T]}$ be defined as Proposition 1. Let $\boldsymbol{J}_{[T]}$ be a series of random sets each i.i.d. sampled from $[N]$ uniformly without replacement with size $N-1$, and let $\mathcal{B}$ be another stochastic iterative algorithm with update rule $\boldsymbol{W}_t = \tilde{\mathcal{M}}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}_{\boldsymbol{J}_t}, \boldsymbol{V}_t, \boldsymbol{J}_t)$ and the same initialization distribution $\mathcal{W}_0$. Given $\boldsymbol{S}$, denote the joint distributions of $(\boldsymbol{W}_{[T]}, \boldsymbol{J}_{[T]}, \boldsymbol{V}_{[T]})$ of $\mathcal{A}$ and $\mathcal{B}$ as $Q^{\boldsymbol{S}}$ and $P^{\boldsymbol{S}}$, respectively. Then, the generalization error of $\mathcal{A}$ can be bounded as:*

$$\mathbb{E}_{\boldsymbol{S},\boldsymbol{V}_{[T]}}\left[\mathcal{R}_{\mathcal{D}}\left(Q_T^{\boldsymbol{S}}(\boldsymbol{W}_T)\right)-\hat{\mathcal{R}}_{\boldsymbol{S}}\left(Q_T^{\boldsymbol{S}}(\boldsymbol{W}_T)\right)\right] \leq \mathbb{E}_{\boldsymbol{S}}\sqrt{\frac{(a_2-a_1)^2}{2}\,\mathrm{KL}\,(Q^{\boldsymbol{S}}||P^{\boldsymbol{S}})}. \tag{6}$$

Theorem 1 is obtained by observing that for fixed $\boldsymbol{V}_{[T]}$ and $\boldsymbol{S}$, $\mathbb{E}_{\boldsymbol{J}}\mathbb{E}_{Q_{s-1}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}}\mathrm{KL}(Q_{s|(s-1)}^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\,\|P_{s|(s-1)}^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}})$ can be viewed as the KL divergence between the joint distribution of $(\mathcal{M}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}, \boldsymbol{V}_t), \boldsymbol{J}_t)$ and $(\tilde{\mathcal{M}}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}_{\boldsymbol{J}_t}, \boldsymbol{V}_t, \boldsymbol{J}_t), \boldsymbol{J}_t)$, where $\boldsymbol{J}_t$ is a random set with size $N-1$ sampled uniformly without replacement from $[N]$, and for fixed $\boldsymbol{J}_t$, $\tilde{\mathcal{M}}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}_{\boldsymbol{J}_t}, \boldsymbol{V}_t, \boldsymbol{J}_t) \sim P_{s|(s-1)}^{\boldsymbol{S}_{\boldsymbol{J}_t},\boldsymbol{V}_{[T]}}$. The detailed proof is deferred to Appendix C.2.

We further introduce an assumption to make Theorem 1 applicable to noise structure analysis.

Assumption 1. *For fixed $P_{1:T}^{\boldsymbol{S}}$ and fixed $\boldsymbol{S}$, under Constraint 1 $(\mathbf{\Sigma}_t(\boldsymbol{S},\cdot))_{t=1}^{T}$ optimizes* KL $\left(Q_{1:T}^{\boldsymbol{S}}||P_{1:T}^{\boldsymbol{S}}\right)$ *when $(\mathbf{\Sigma}_t(\boldsymbol{S},\cdot))_{t=1}^{T}$ optimizes* KL $\left(P_{1:T}^{\boldsymbol{S}}||Q_{1:T}^{\boldsymbol{S}}\right)$.

By combining Theorem 1 and Assumption 1, we only need to optimize $(P^{\boldsymbol{S}}_{1:T}||Q^{\boldsymbol{S}}_{1:T})$ with respect to $\boldsymbol{\Sigma}_{[T]}(\boldsymbol{S},\cdot)$ subject to Constraint 1, which can be further decomposed into $\sum_{s=1}^{T}\mathbb{E}_{P^{\boldsymbol{S}}_{1:s-1}}$ $\mathrm{KL}(P^{\boldsymbol{S}}_{s|(s-1)}\|Q^{\boldsymbol{S}}_{s|(s-1)})$. In this decomposition, the distribution $P^{\boldsymbol{S}}_{1:s}$ is independent of $\boldsymbol{\Sigma}_t$ for any $s,t\in[T]$ and hence only each KL term $\mathrm{KL}\left(P^{\boldsymbol{S}}_{s|(s-1)}\|Q^{\boldsymbol{S}}_{s|(s-1)}\right)$ depends on $\boldsymbol{\Sigma}_s$. We only need to optimize $\mathrm{KL}\left(P^{\boldsymbol{S}}_{s|(s-1)}\|Q^{\boldsymbol{S}}_{s|(s-1)}\right)$ with respect to $\boldsymbol{\Sigma}_s(\boldsymbol{S},\cdot)$, which is clear and easy to deal with.

However, Proposition 1 can NOT be used to analyze the noise covariance even with a similar assumption. Since $\mathrm{KL}(Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}_{s|(s-1)}\|P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s|(s-1)})$ is dependent on $\boldsymbol{\Sigma}_s$ for any $\boldsymbol{J}$, we need to assume $\mathbb{E}_{\boldsymbol{J}}\mathbb{E}_{\boldsymbol{V}_{[T]}}\,\mathrm{KL}(Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\|P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s|(s-1)})$ and $\mathbb{E}_{\boldsymbol{J}}\mathbb{E}_{\boldsymbol{V}_{[T]}}\,\mathrm{KL}(P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s|(s-1)}\|Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}_{s|(s-1)})$ are simultaneously optimized, and optimize $\mathbb{E}_{\boldsymbol{S}}\sqrt{\mathrm{KL}\left((Q^{\boldsymbol{S}},\boldsymbol{J},\boldsymbol{V}_{[T]})||(P^{\boldsymbol{S}},\boldsymbol{J},\boldsymbol{V}_{[T]})\right)}$ instead. However, this is exactly (a weaker version of) Theorem 2, which can not tackle **Problem 1** by the discussion below Theorem 2.

On the other hand, Theorem 1 and Assumption 1 is not suitable to solve **Problem 2** due to $\mathcal{B}$ takes the whole data set $\boldsymbol{S}$ as input. To tackle **Problem 2**, we provide another generalization bound which can be viewed as Proposition 1 with reversed order of prior and posterior.

Theorem 2. *Let $\boldsymbol{S}$, $\boldsymbol{V}_{[T]}$, $\mathcal{A}$ and $\mathcal{B}$ be defined as Proposition 1. Then, the generalization error of $\mathcal{A}$ can be bounded as*

$$\mathrm{Gerr}\leq\mathbb{E}_{\boldsymbol{S},\boldsymbol{J},\boldsymbol{V}_{[T]}}\sqrt{\frac{(a_2-a_1)^2}{2}\,\mathrm{KL}\left(P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}\,\middle\|Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}\right)}. \tag{7}$$

The reason that Theorem 2 enables us to optimize the generalization error bound with respect to the noise covariance under Constraint 1 is two-fold. First, the order of $Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}$ and $P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}$ in Eq. (7) is reversed compared to Proposition 1. Then, by Lemma 1, the KL term in Eq. (7) can be further decomposed into $\sum_{s=1}^{T}\mathbb{E}_{P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{1:s-1}}\,\mathrm{KL}(P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s|(s-1)}||Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}_{s|(s-1)})$, in which only each KL term depends on each $\boldsymbol{\Sigma}_s$, respectively. Secondly, for each summand in the above decomposition, only $Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}_{s|(s-1)}$ depends on $\boldsymbol{S}_{\boldsymbol{J}^c}$, making it easy to compute the optimal prior in a "greedy" sense (see details in Section 6).

On the other hand, Theorem 2 is not suitable to analyze **Problem 1** because $\mathbb{E}_{P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s-1}}\,\mathrm{KL}(P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s|(s-1)}||Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}_{s|(s-1)})$ will depend on $\boldsymbol{\Sigma}_s(\boldsymbol{S},\cdot)$ for any $\boldsymbol{J}$, and we will need to (approximately) optimize $\mathbb{E}_{\boldsymbol{J}}\mathbb{E}_{P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s-1}}\,\mathrm{KL}(P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s|(s-1)}||Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}_{s|(s-1)})$, which is equivalent to optimizing $\mathbb{E}_{\boldsymbol{J}}P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s-1}(\boldsymbol{W}_{s-1})\,\mathrm{KL}(P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s|(s-1)}||Q^{\boldsymbol{S},\boldsymbol{V}_{[T]}}_{s|(s-1)})$ with respect to $\boldsymbol{\Sigma}_s(\boldsymbol{S},\boldsymbol{W}_{s-1})$. Since the exact form of $P^{\boldsymbol{S}_{\boldsymbol{J}},\boldsymbol{V}_{[T]}}_{s-1}$ can be very complicated, analyzing **Problem 1** by Theorem 2 is not proper.

5 Solve Problem 1: Fixed Data-dependent Prior

In this section, we optimize the generalization bound of state-dependent SGLD with respect to the noise structure for a fixed commonly-used data-dependent prior. Specifically, by Theorem 1 and Assumption 1, the generalization bound at T is given by

$$\mathrm{Gen}_T=\mathbb{E}_{\boldsymbol{S}}\sqrt{\mathrm{KL}\left(Q^{\boldsymbol{S}}_{1:T}||P^{\boldsymbol{S}}_{1:T}\right)}, \tag{8}$$

where $Q^{\boldsymbol{S}}_{1:T}$ is the joint distribution of $(\boldsymbol{W}_{[T]},\boldsymbol{J}_{[T]},\boldsymbol{V}_{[T]})$ defined by Theorem 1 with $\mathcal{M}(\boldsymbol{W}_{t-1},\boldsymbol{S},\boldsymbol{V}_t)$ being the update rule of state-dependent SGLD (Eq. (3)). With a fixed data-dependent prior, Gen_T is then a functional of noise covariance $\boldsymbol{\Sigma}_{[T]}$, where we use $\mathrm{Gen}_T(\boldsymbol{\Sigma}_{[T]})$ to denote the dependency. Our goal is to find the solution $\boldsymbol{\Sigma}^*_{[T]}$ for the following optimization problem :

$$\text{Minimize: }\ \mathrm{Gen}_T(\boldsymbol{\Sigma}_{[T]})\qquad\text{Subject to: Constraint 1.} \tag{9}$$

It requires selecting the data-dependent prior in Theorem 1 to calculate the generalization error. For fair comparison, we select the data-dependent prior as in [24], which (with the notations in Theorem

1) is given by

$$\tilde{\mathcal{M}}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}_{\boldsymbol{J}_t}, \boldsymbol{V}_t, \boldsymbol{J}_t)$$
$$= \boldsymbol{W}_{t-1} - \eta_t \left(\frac{|\boldsymbol{V}_t \cap \boldsymbol{J}_t|}{|\boldsymbol{V}_t|} \nabla \mathcal{R}_{\boldsymbol{S}_{\boldsymbol{V}_t \cap \boldsymbol{J}_t}} (\boldsymbol{W}_{t-1}) + \frac{|\boldsymbol{V}_t \cap \boldsymbol{J}_t^c|}{|\boldsymbol{V}_t|} \nabla \mathcal{R}_{\boldsymbol{S}_{\boldsymbol{J}_t}} (\boldsymbol{W}_{t-1}) \right) + \mathcal{N}(\boldsymbol{0}, \sigma_t \mathbb{I}_d), \quad (10)$$

where $\boldsymbol{V}_t$ is the same as that in Eq. (3), and $\sigma_t > 0$ is the noise scale of prior noise covariance. We further set $c_t(\boldsymbol{S}, \boldsymbol{W}) \leq \sigma_t$ in order to guarantee that the noise scale of the prior is comparable to that of the posterior. We defer the formal description of the iteration of the prior to Appendix D.1.

By Eq. (10), we obtain the KL divergence between $(\tilde{\mathcal{M}}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}_{\boldsymbol{J}_t}, \boldsymbol{V}_t, \boldsymbol{J}_t), \boldsymbol{J}_t, \boldsymbol{V}_t)$ and $(\mathcal{M}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}, \boldsymbol{V}_t), \boldsymbol{J}_t, \boldsymbol{V}_t)$ as the following Lemma:

Lemma 3. *Let $\tilde{\mathcal{M}}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}_{\boldsymbol{J}_t}, \boldsymbol{V}_t, \boldsymbol{J}_t)$ and $\mathcal{M}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}, \boldsymbol{V}_t)$ be defined respectively as Eq.(10) and Eq.(3). Then, for fixed $\boldsymbol{W}_{t-1}$, the KL divergence between $(\tilde{\mathcal{M}}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}_{\boldsymbol{J}_t}, \boldsymbol{V}_t, \boldsymbol{J}_t), \boldsymbol{J}_t, \boldsymbol{V}_t)$ and $(\mathcal{M}_t(\boldsymbol{W}_{t-1}, \boldsymbol{S}, \boldsymbol{V}_t), \boldsymbol{J}_t, \boldsymbol{V}_t)$ can be calculated as*

$$\mathrm{KL}\left(P^{\boldsymbol{S}}_{t|(t-1)} \middle\| Q^{\boldsymbol{S}}_{t|(t-1)} \right) = \frac{1}{2} \left(\sigma_t \operatorname{tr}\left(\boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1})^{-1} \right) + \ln\left(\det \boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1}) \right) - d \ln 2\sigma_t - d \right)$$
$$+ \frac{\eta_t^2}{2Nb_t} \left(\frac{N}{N-1} \right)^2 \operatorname{tr}\left(\boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W}_{t-1})^{-1} \boldsymbol{\Sigma}^{sd}_{\boldsymbol{S}, \boldsymbol{W}_{t-1}} \right).$$

This lemma allows us to obtain the KL divergence along the full trajectory. We defer the proof of Lemma 3 to Appendix D.2. With the data-dependent prior defined as Eq.(10), we are now ready to solve the optimization problem (9). Specifically, we first show how the optimal point looks like.

Theorem 3. *Let prior and posterior be defined as Eq.(10) and Eq.(3), respectively. Then, the solution of Problem 9 is given by*

$$\boldsymbol{\Sigma}_t^*(\boldsymbol{S}, \boldsymbol{W}) = \boldsymbol{Q}^{sd}_{\boldsymbol{S}, \boldsymbol{W}} \operatorname{Diag}(\tilde{\sigma}_1^{\boldsymbol{S}, \boldsymbol{W}}, \cdots, \tilde{\sigma}_d^{\boldsymbol{S}, \boldsymbol{W}}) \left(\boldsymbol{Q}^{sd}_{\boldsymbol{S}, \boldsymbol{W}} \right)^\top,$$

where $\tilde{\sigma}_i^{\boldsymbol{S}, \boldsymbol{W}} \geq 0$ $(i \in [d])$ (the exact form is omitted here) and $\left(\boldsymbol{Q}^{sd}_{\boldsymbol{S}, \boldsymbol{W}} \right)$ is the orthogonal matrix that diagonalizes $\boldsymbol{\Sigma}^{sd}_{\boldsymbol{S}, \boldsymbol{W}}$ as

$$\boldsymbol{\Sigma}^{sd}_{\boldsymbol{S}, \boldsymbol{W}} = \boldsymbol{Q}^{sd}_{\boldsymbol{S}, \boldsymbol{W}} \operatorname{Diag}(\sigma_1^{\boldsymbol{S}, \boldsymbol{W}}, \cdots, \sigma_d^{\boldsymbol{S}, \boldsymbol{W}}) \left(\boldsymbol{Q}^{sd}_{\boldsymbol{S}, \boldsymbol{W}} \right)^\top.$$

Moreover, for any $i \neq j$, $\tilde{\sigma}_i^{\boldsymbol{S}, \boldsymbol{W}} \geq \tilde{\sigma}_j^{\boldsymbol{S}, \boldsymbol{W}}$ if and only if $\sigma_i^{\boldsymbol{S}, \boldsymbol{W}} \geq \sigma_j^{\boldsymbol{S}, \boldsymbol{W}}$.

From Theorem 3, the optimal point $(\boldsymbol{\Sigma}_t^*)_{t=1}^T$ is similar to the empirical gradient covariance matrix $\boldsymbol{\Sigma}^{sd}$: 1) $(\boldsymbol{\Sigma}_t^*)_{t=1}^T$ share the same eigenvectors with $\boldsymbol{\Sigma}^{sd}$; 2) the eigenvalues of $\boldsymbol{\Sigma}_t^*$ has the same order as $\boldsymbol{\Sigma}^{sd}$. We defer the exact formula of $\tilde{\sigma}_i^{\boldsymbol{S}, \boldsymbol{W}}$ in Theorem 3 to Appendix D.4.

On the other hand, $(\tilde{\sigma}_i^{\boldsymbol{S}, \boldsymbol{W}})_{i=1}^d$ does not exactly match (scaled) $(\sigma_i^{\boldsymbol{S}, \boldsymbol{W}})_{i=1}^d$. The major reason is the isotropy of the gradient noise of the data-dependent prior. As indicated by Lemma 3, the value of Gen_T is affected by two factors: 1) the distance between means of the prior and the posterior; 2) the distance between covariances of the prior and the posterior. Therefore, the exact form of the eigenvalues of $\boldsymbol{\Sigma}_t^*(\boldsymbol{S}, \boldsymbol{W})$ behaves like a "mid-point" of the isotropic covariance of the prior and the empirical gradient covariance under the effect of these two factors.

We next provide **a proof sketch of Theorem 3**. The main idea is by applying the decomposition of the KL divergence and optimizing each step's KL divergence separately. By Lemma 1, we have

$$\mathrm{Gen}_T = \mathbb{E}_{\boldsymbol{S}} \sqrt{\sum_{s=1}^T \mathbb{E}_{P^{\boldsymbol{S}}_{s-1}} \mathrm{KL}\left(P^{\boldsymbol{S}}_{s|(s-1)} \middle\| Q^{\boldsymbol{S}}_{s|(s-1)} \right)}.$$

Because $P^{\boldsymbol{S}}_{1:s}$ is independent of $\boldsymbol{\Sigma}_t$ for any $s, t \in [T]$ as we argued in Section 4.1, we only need to optimize $\mathrm{KL}\left(P^{\boldsymbol{S}}_{s|(s-1)} \middle\| Q^{\boldsymbol{S}}_{s|(s-1)} \right)$ that is characterized by Lemma 3 with respect to $\boldsymbol{\Sigma}_s$ separately. We introduce the following lemma to characterize the optimal $\boldsymbol{\Sigma}_s(\boldsymbol{S}, \boldsymbol{W})$ for any given $\boldsymbol{S}$ and $\boldsymbol{W}$.

Lemma 4. *Let $\boldsymbol{B} \in \mathbb{R}^{d \times d}$ be a (fixed) positive definite matrix with eigenvalues $(\beta_1, \cdots, \beta_d)$, where $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_d > 0$ are positive reals. Let $\boldsymbol{G} \in \mathbb{R}^{d \times d}$ be a positive definite matrix variable with $\operatorname{tr} \boldsymbol{G} = c$, where c is a positive constant and $c \leq \operatorname{tr}(\boldsymbol{B})$. Then the minimum of*

$\operatorname{tr}(\boldsymbol{G}^{-1}\boldsymbol{B}) + \ln\det(\boldsymbol{G})$ *is achieved at* $\boldsymbol{G} = \boldsymbol{P}^\top \operatorname{Diag}(\alpha_1, \cdots, \alpha_d)\boldsymbol{P}$*, where* $\alpha_{[d]}$ *are some positive reals which satisfy* $\beta_i \geq \beta_j$ *if and only if* $\alpha_i \geq \alpha_j$ *for any* $i, j \in [d]$*, and* $\boldsymbol{P}$ *is any orthogonal matrix which diagonalizes* $\boldsymbol{B}$ *as* $\boldsymbol{B} = \boldsymbol{P}^\top \operatorname{Diag}(\beta_1, \cdots, \beta_d)\boldsymbol{P}$*.*

Lemma 4 is proved via two steps: 1) we first prove for $\boldsymbol{G}$ with fixed eigenvalues, $\operatorname{tr}\boldsymbol{G}^{-1}\boldsymbol{B} + \ln\det(\boldsymbol{G})$ is optimized if and only if $\boldsymbol{G}$ and $\boldsymbol{B}$ share the same eigenvectors; 2) we then calculate the eigenvalues of the optimal $\boldsymbol{G}$ using the method of Lagrange multipliers. Theorem 3 can then be obtained by applying Lemma 4 and setting $\boldsymbol{G} = \boldsymbol{\Sigma}_t(\boldsymbol{S}, \boldsymbol{W})$ and $\boldsymbol{B} = \sigma_t \mathbb{I} + \frac{\eta_t^2}{N b_t}\left(\frac{N}{N-1}\right)^2 \boldsymbol{\Sigma}^{sd}_{\boldsymbol{S},\boldsymbol{W}}$.

6 Solve Problem 2 with Optimal Prior

In this section, we solve **Problem 2** through Theorem 2 with greedily selected priors. Similar to Section 5, we optimize the generalization bound in Theorem 2 under Constraint 1. To simplify the calculation, we optimize a slightly weakened version of the bound in Theorem 2 as

$$\widetilde{\operatorname{Gen}}_T = \mathbb{E}_{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}, \boldsymbol{J}} \sqrt{\frac{(a_2 - a_1)^2}{2} \mathbb{E}_{\boldsymbol{S_{J^c}} \sim \mathcal{D}} \operatorname{KL}\left(P^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}} \middle\| Q^{\boldsymbol{S}, \boldsymbol{V}_{[T]}}\right)}, \tag{11}$$

which can be obtained by applying Jessen's Inequality to Theorem 2. Also, we further restrict the considered noise covariance $\boldsymbol{\Sigma}_{[T]}$ only depending on $\boldsymbol{W}$, and independent of sample $\boldsymbol{S}$. By Lemma 1, the KL term in $\widetilde{\operatorname{Gen}}_T$ can be decomposed into

$$\mathbb{E}_{\boldsymbol{S_{J^c}} \sim \mathcal{D}} \operatorname{KL}\left(P^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}} || Q^{\boldsymbol{S}, \boldsymbol{V}_{[T]}}\right) = \sum_{s=1}^{T} \mathbb{E}_{\boldsymbol{z}_N \sim \mathcal{D}} \mathbb{E}_{P_{s-1}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}} \operatorname{KL}\left(P_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}} || Q_{s|(s-1)}^{\boldsymbol{S}, \boldsymbol{V}_{[T]}}\right). \tag{12}$$

Therefore, when optimizing $\widetilde{\operatorname{Gen}}_T$ with respect to prior P, both $\operatorname{KL}(P_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}} || Q_{s|(s-1)}^{\boldsymbol{S}, \boldsymbol{V}_{[T]}})$ and $P_{i-1}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}$ $(i > s)$ have dependence on $P_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}$. Similar to the discussion in Section 4.1, the dependence of $P_{i-1}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}$ with $i > s$ on $P_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}$ can be very complex. Therefore, for any fixed $\boldsymbol{\Sigma}_{[T]}$, we approximate the optimal $P_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}$ by using instead the optimal of only the s-th term in the decomposition Eq.(12):

$$P^{*}{}_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}} = \arg\min_{P \in \Delta(\mathbb{R}^d)} \mathbb{E}_{\boldsymbol{J}^c \sim \mathcal{D}} \operatorname{KL}\left(P \middle\| Q_{s|(s-1)}^{\boldsymbol{S}, \boldsymbol{V}_{[T]}}\right). \tag{13}$$

Here $P^{*}{}_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}$ is dependent on $\boldsymbol{\Sigma}_{[T]}$. We call $(P^{*}{}_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}})_{s=1}^{T}$ the *greedy prior*, in the sense that $P^{*}{}_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}$ optimizes $\widetilde{\operatorname{Gen}}_s$ as demonstrated by the following lemma:

Lemma 5. *For fixed* $\boldsymbol{\Sigma}_{[T]}$*,* $P^{*}{}_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}$ *defined by Eq.(13) optimizes* $\widetilde{\operatorname{Gen}}_s$*.*

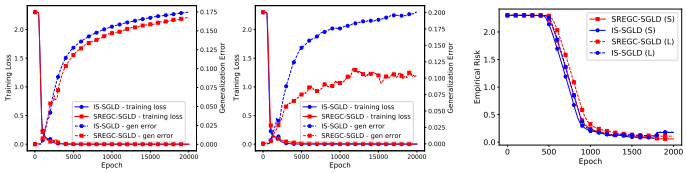
The proof of Lemma 5 can be found in Appendix E.1. By Lemma 5, $P^{*}{}_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}$ is the **optimal** prior to select if we are optimizing the generalization error of the first s steps. Therefore, when optimizing the generalization error bound Eq.(11) iteratively, it is natural to use $P^{*}{}_{s|(s-1)}^{\boldsymbol{S_J}, \boldsymbol{V}_{[T]}}$ as the prior which optimizes the generalization error bound for the current parameters. Then, the optimal posterior with greedily-chosen prior is characterized by the following theorem:

Theorem 4. *Let the iteration of SGLD with state-dependent noise* $Q^{\boldsymbol{S}, \boldsymbol{V}_{[T]}}$ *be given as Eq.(3). For any fixed* $\boldsymbol{\Sigma}_{[T]}$*, the greedy optimal prior of step* t *is given by*

$$\boldsymbol{W}_t = \boldsymbol{W}_{t-1} - \eta_t \left(\frac{|\boldsymbol{V}_t \cap \boldsymbol{J}|}{|\boldsymbol{V}_t|} \nabla \mathcal{R}_{\boldsymbol{S}_{\boldsymbol{V}_t \cap \boldsymbol{J}}}(\boldsymbol{W}_{t-1}) + \frac{|\boldsymbol{V}_t \cap \boldsymbol{J}^c|}{|\boldsymbol{V}_t|} \nabla \mathcal{R}_{\mathcal{D}}(\boldsymbol{W}_{t-1}) \right) + \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t(\boldsymbol{W}_{t-1})).$$

Furthermore, under Constraint 1, the optimal covariance of noise $\boldsymbol{\Sigma}^*_{[T]}$ *for* $\widetilde{\operatorname{Gen}}_T$ *with the greedy optimal prior is given by* $\boldsymbol{\Sigma}^*_t(\boldsymbol{W}) = \lambda_t(\boldsymbol{W})\,(\boldsymbol{\Sigma}^{pop}_{\boldsymbol{W}})^{\frac{1}{2}}$*, where* $\lambda_t(\boldsymbol{W}) = c_t(\boldsymbol{W}) / \operatorname{tr}((\boldsymbol{\Sigma}^{pop}_{\boldsymbol{W}})^{\frac{1}{2}})$*.*

The proof of Theorem 4 requires a standard analysis of mixture of logarithmic Gaussian probability density together with applying Lemma 4, which is deferred to Appendix E.2. By law of large numbers, we have $\boldsymbol{\Sigma}^{sd}_{\boldsymbol{S},\boldsymbol{W}} \approx \boldsymbol{\Sigma}^{pop}_{\boldsymbol{W}}$ as sample size N is large enough, which demonstrates the similarity between the solutions of **Problem 1** and **Problem 2**. Also, as N is large enough, $\nabla \mathcal{R}_{\boldsymbol{S_J}} \approx \nabla \mathcal{R}_{\mathcal{D}}$, and the mean of the greedy optimal prior recovers the mean of the prior used in [24] (Eq. 10).



(a) trace scale $= 1$ (b) trace scale $= 5$ (c) All training loss

Figure 2: The training loss and and the generalization error (refer to Eq. (1), test loss $-$ training loss) of the IS-SGLD and SREGC-SGLD. (a) and (b) correspond to noise covariance trace scale 1 and 5, respectively. Training loss in (a) and (b) is plot together in (c) to compare the convergence behavior.

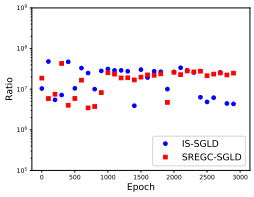


Figure 3: Ratio of the 1st to the 500th largest eigenvalue of the empirical gradient covariance for two SGLDs.

7 Empirical Verification

In this section, we support our theoretical findings with some experiments. We adopt the setting in [37, Fig.3]: using a four-layer neural network to conduct the classification task on the Fashion-MNIST except that we use 10000 training samples instead of 1200 in [37]. We defer detailed settings of the experiments to Appendix F. We apply SGLD with different noise covariances to optimize the above task. Specifically, one SGLD adds noise with covariance the same as the (scaled) square root of the empirical gradient covariance, which is called "SREGC-SGLD". The other SGLD adds isotropic noise, which is called "IS-SGLD". Moreover, the IS-SGLD's noise is always scaled to have the same trace as that of SREGC-SGLD, in order to satisify the Constraint 1.

7.1 Validation of Generalization Performance

In this subsection, we verify the claim of Theorem 4 by comparing the generalization error between IS-SGLD and SREGC-SGLD. Specifically, we plot the training loss and and the generalization error (test loss $-$ training loss) of the two SGLDs across iterations with different scales in Fig.2. From Fig.2, we can see in Fig.2a with the trace of the noise covariance $\mathrm{tr}((\boldsymbol{\Sigma}^{sd}_{\boldsymbol{S},\boldsymbol{W}})^{1/2})$ and Fig.2b with the trace of the noise covariance $5\,\mathrm{tr}((\boldsymbol{\Sigma}^{sd}_{\boldsymbol{S},\boldsymbol{W}})^{1/2})$, the training loss behaves similar for two SGLDs with different noise structures, which supports Lemma 2. Secondly, we can see from Fig.2a and Fig.2b that the generalization error of SREGC-SGLD is smaller than IS-SGLD, which supports Theorem 4. Finally, we can see from Fig.2c that SGLDs with trace $\mathrm{tr}((\boldsymbol{\Sigma}^{sd}_{\boldsymbol{S},\boldsymbol{W}})^{1/2})$ has a faster convergent rate than those with trace $5\,\mathrm{tr}((\boldsymbol{\Sigma}^{sd}_{\boldsymbol{S},\boldsymbol{W}})^{1/2})$, which supports Lemma 2.

7.2 How Anisotropic Is the Empirical Gradient Covariance?

In Section 5 and Section 6, we have shown that both the optimal noise covariances of two problems are similar to the (scaled) empirical gradient covariance. It is important to study to what extent the empirical gradient covariance is anisotropic. We plot the ratio of the largest eigenvalue to the 500th largest eigenvalue of the empirical gradient covariance along the training trajectory of both SREGC-SGLD and IS-SGLD in Figure 3. By Figure 3, we can see that throughout the training procedure, the ratios stay around 10^7 for both SGLDs, which demonstrates that the energy of empirical gradient covariance concentrates in a very small subspace (less than 5% of the total 11330 parameter dimensions). Hence the empirical gradient covariance is highly anisotropic.

8 Conclusion and Future Direction

In this paper, we provide a new way to improve the information theoretical generalization: optimizing the bound with respect to the noise structure of SGLD under a constraint on the trace of the noise covariance. We propose two problems that differ by whether the prior is optimized or given and develop new information theoretical bounds to solve them. Interestingly, we observe that the optimal noise covaraince aligns with the empirical gradient covariance. One limitation, also an interesting problem, is that we solve Problem 1 based on the Assumption 1. It is intriguing if one can prove the reverse-ordered KL for the case of Theorem 1, which can automatically remove the Assumption 1 and solve Problem 1.

References

[1] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

[2] A. R. Asadi, E. Abbe, and S. Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[3] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[4] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer, 2003.

[5] Y. Bu, S. Zou, and V. V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130, 2020.

[6] Z. Charles and D. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754. PMLR, 2018.

[7] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, 2001.

[8] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, 2015.

[9] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[10] V. Feldman and T. Steinke. Calibrating noise to variance in adaptive data analysis. In *Conference On Learning Theory*, pages 535–544. PMLR, 2018.

[11] S. B. Gelfand and S. K. Mitter. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

[12] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[13] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *arXiv preprint arXiv:2004.12983*, 2020.

[14] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[15] I. Issa, A. R. Esposito, and M. Gastpar. Strengthened information-theoretic bounds on the generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 582–586. IEEE, 2019.

[16] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

[17] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[18] I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.

[19] Y. Lei and Y. Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2021.

[20] J. Li, X. Luo, and M. Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2019.

[21] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

[22] Q. Meng, S. Gong, W. Chen, Z.-M. Ma, and T.-Y. Liu. Dynamic of stochastic gradient descent with state-dependent noise. *arXiv preprint arXiv:2006.13719*, 2020.

[23] W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638. PMLR, 2018.

[24] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. *arXiv preprint arXiv:1911.02151*, 2019.

[25] B. Neyshabur, S. Bhojanapalli, and N. Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

[26] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. Pac-bayes bounds with data dependent priors. *The Journal of Machine Learning Research*, 13(1):3507–3531, 2012.

[27] A. Pensia, V. Jog, and P.-L. Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550. IEEE, 2018.

[28] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

[29] D. Russo and J. Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240. PMLR, 2016.

[30] T. Steinke and L. Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

[31] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[32] Y. Wen, K. Luk, M. Gazeau, G. Zhang, H. Chan, and J. Ba. An empirical study of large-batch stochastic gradient descent with structured covariance noise. *arXiv e-prints*, pages arXiv–1902, 2019.

[33] Z. Xie, I. Sato, and M. Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast. *arXiv preprint arXiv:2002.03495*, 2020.

[34] A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *arXiv preprint arXiv:1705.07809*, 2017.

[35] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[36] W. Zhou, V. Veitch, M. Austern, R. P. Adams, and P. Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. In *International Conference on Learning Representations*, 2018.

[37] Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] , please refer to Section 8.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] , please refer to Assumption 1.
   (b) Did you include complete proofs of all theoretical results? [Yes] , please refer to the proof skeleton under Theorems 1, 2, 3, and 4. Detailed proofs are deferred to the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] , please refer to Section 7 and Appendix F.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] , please refer to Appendix F.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] , Fashion-MNIST is public.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]