Asymptotically Optimal Generalization Error Bounds for Noisy, Iterative Algorithms

Ibrahim Issa IBRAHIM.ISSA@AUB.EDU.LB
American University of Beirut, Lebanon, and École Polytechnique Fédérale de Lausanne, Switzerland

Amedeo Roberto Esposito AMEDEOROBERTO.ESPOSITO@IST.AC.AT
Institute of Science and Technology Austria

Michael Gastpar MICHAEL.GASTPAR@EPFL.CH
École Polytechnique Fédérale de Lausanne, Switzerland

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We adopt an information-theoretic framework to analyze the generalization behavior of the class of iterative, noisy learning algorithms. This class is particularly suitable for study under information-theoretic metrics as the algorithms are inherently randomized, and it includes commonly used algorithms such as Stochastic Gradient Langevin Dynamics (SGLD). Herein, we use the maximal leakage (equivalently, the Sibson mutual information of order infinity) metric, as it is simple to analyze, and it implies both bounds on the probability of having a large generalization error and on its expected value. We show that, if the update function (e.g., gradient) is bounded in L_2-norm, then adding isotropic Gaussian noise leads to optimal generalization bounds: indeed, the input and output of the learning algorithm in this case are asymptotically statistically independent. Furthermore, we demonstrate how the assumptions on the update function affect the optimal (in the sense of minimizing the induced maximal leakage) choice of the noise. Finally, we compute explicit tight upper bounds on the induced maximal leakage for several scenarios of interest.

Keywords: Noisy iterative algorithms, generalization error, maximal leakage, Gaussian noise

1. Introduction

One of the key challenges in machine learning research concerns the “generalization” behavior of learning algorithms. That is: if a learning algorithm performs well on the training set, what guarantees can one provide on its performance on new samples?

While the question of generalization is understood in many settings (Bousquet et al., 2003; Shalev-Shwartz and Ben-David., 2014), existing bounds and techniques provide vacuous expressions when employed to show the generalization capabilities of deep neural networks (DNNs) (Bartlett et al., 2017, 2019; Jiang et al., 2020; Zhang et al., 2021). In general, classical measures of model expressivity (such as Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1991), Rademacher complexity (Bartlett and Mendelson, 2003), etc.) fail to explain the generalization abilities of DNNs due to the fact that they are typically over-parameterized models with less training data than model parameters. A novel approach was introduced by (Russo and Zou, 2016), and (Xu and Raginsky, 2017) (further developed by Steinke and Zakynthinou (2020); Bu et al. (2020); Esposito et al. (2021); Esposito and Gastpar (2022) and many others), where information-theoretic techniques are used to link the generalization capabilities of a learning algorithm to information measures. These quantities are algorithm-dependent and can be used to analyze the generalization capabilities of general classes of updates and models *e.g.*, noisy iterative algorithms such as the Stochastic Gradient Langevin Dynamics (SGLD) (Pensia et al., 2018; Wang et al., 2021), which can thus be applied to deep learning settings. Moreover, it has been shown that information-theoretic bounds can be non-vacuous and reflect the real generalization behavior even in deep learning settings (Dziugaite and Roy, 2017; Zhou et al., 2018; Negrea et al., 2019; Haghifam et al., 2020).

In this work we adopt and expand the framework introduced by Pensia et al. (2018), but instead of focusing on the mutual information between the input and output of an iterative algorithm, we compute the maximal leakage (Issa et al., 2020). Maximal leakage, together with other information measures of the Sibson/Rényi family (maximal leakage can be shown to be Sibson Mutual information of order infinity (Issa et al., 2020)), have been linked to high-probability bounds on the generalization error (Esposito et al., 2021). In particular, given a learning algorithm $\mathcal{A}$ trained on data-set S (made of n samples), one can provide the following guarantee in the case of the $0-1$ loss:

$$\mathbf{Pr}(|\text{gen-err}(\mathcal{A},S)| \geq \eta) \leq 2\exp(-2n\eta^2 + \mathcal{L}\left(S\!\to\!\mathcal{A}(S)\right)), \tag{1}$$

where $\mathcal{L}\left(S\!\to\!\mathcal{A}(S)\right)$ is defined in equation (2) below. This deviates from much of the literature in which the focus is on bounding the **expected** generalization error instead (Xu and Raginsky, 2017; Steinke and Zakynthinou, 2020). Consequently, if one can guarantee that for a class of algorithms, the maximal leakage between the input and the output is bounded, then one can provide an **exponentially decaying** (in the number of samples n) bound on the probability of having a large generalization error. This is in general not true for mutual information, which can typically only guarantee a linearly decaying bound on the probability of the same event (Bassily et al., 2018). Moreover, a bound on maximal leakage implies a bound on mutual information (cf. Equation (6)) and, consequently, a bound on the expected generalization error of $\mathcal{A}$ (exploiting the link between mutual information and expected generalization error (Xu and Raginsky, 2017)). The main advantage of maximal leakage lies in the fact that it depends on the distribution of the samples only through its support. It is thus naturally independent from the distribution over the samples and particularly amenable to analysis, especially in additive noise settings.

The contributions of this work can be summarized as follows:

- we derive novel bounds on $\mathcal{L}\left(S\!\to\!\mathcal{A}(S)\right)$ whenever $\mathcal{A}$ is a noisy, iterative algorithm (SGLD-like), which then implies the first bounds showing generalization with high-probability of said mechanisms;
- we show that the bounds provided on maximal leakage strictly improve the bounds provided by Pensia et al. (2018), and we thus provide a tighter bound on the expected generalization error of said algorithms as well;
- we show that, under certain assumptions, adding Gaussian noise is asymptotically optimal in the number of dimensions d. In particular, we prove that the maximal leakage (and, consequently, the mutual information) between the input and output of this family of algorithms goes to 0 with the number of dimensions. This implies that the input and output are *asymptotically independent* which is consistent with practical observations: larger neural networks generalize better;
- we leverage the analysis to extrapolate to optimize the type of noise to be added (in the sense that minimizing the induced maximal leakage), based on the assumptions imposed on the algorithm. In particular,

- if one assumes the L_p norm of the gradient to be bounded, with $p \leq 2$, our analysis shows that adding Gaussian noise is asymptotically optimal (as discussed above);
- if one assumes the L_∞ norm of the gradient to be bounded, then adding uniform noise is minimizes the maximal leakage upper bound.

Hence, the analysis and computation of maximal leakage can *also* be used to inform the design of novel noisy, iterative algorithms.

1.1. Related Work

The line of work exploiting information measures to bound the expected generalization started in (Russo and Zou, 2016; Xu and Raginsky, 2017) and was then refined with a variety of approaches considering Conditional Mutual Information (Steinke and Zakynthinou, 2020; Haghifam et al., 2020), the Mutual Information between individual samples and the hypothesis (Bu et al., 2019) or improved versions of the original bounds (Issa et al., 2019; Hafez-Kolahi et al., 2020). Other approaches employed the Kullback-Leibler Divergence with a PAC-Bayesian approach (McAllester, 2013; Zhou et al., 2018). Moreover, said bounds were then characterized for specific SGLD-like algorithms, denoted as "noisy, iterative algorithms" and used to provide novel, non-vacuous bounds for Neural Networks (Pensia et al., 2018; Negrea et al., 2019; Haghifam et al., 2020; Wang et al., 2023) as well as for SGD algorithms (Neu et al., 2021). Recent efforts tried to provide the optimal type of noise to add in said algorithms and reduce the (empirical) gap in performance between SGLD and SGD (Wang et al., 2021). All of these approaches considered the KL-Divergence or (variants of) Shannon's Mutual Information. General bounds on the expected generalization error leveraging arbitrary divergences were given in (Esposito and Gastpar, 2022; Lugosi and Neu, 2022). Another line of work considered instead bounds on the probability of having a large generalization error (Bassily et al., 2018; Esposito et al., 2021; Hellström and Durisi, 2020) and focused on large families of divergences and generalizations of the Mutual Information (in particular of the Sibson/Rényi-family, including conditional versions).

2. Preliminaries, Setup, and a General Bound

2.1. Preliminaries

2.1.1. Information Measures

The main building block of the information measures considered in this work is the Rényi's α-divergence between two measures P and Q, $D_\alpha(P\|Q)$ (which can be seen as a parametrized generalization of the Kullback Leibler-divergence) (van Erven and Harremoës, 2014, Definition 2). Starting from Rényi's Divergence and the geometric averaging that it involves, Sibson built the notion of Information Radius (Sibson, 1969) which can be seen as a special case of the following quantity (Verdú, 2015): $I_\alpha(X,Y) = \min_{Q_Y} D_\alpha(P_{XY}\|P_X Q_Y)$. Sibson's $I_\alpha(X,Y)$ represents a generalization of Shannon's mutual information, indeed one has that: $\lim_{\alpha\to 1} I_\alpha(X,Y) = I(X;Y) = \mathbb{E}_{P_{XY}}\left[\log\left(\frac{dP_{XY}}{dP_X P_Y}\right)\right]$. Differently, when $\alpha \to \infty$, one gets:

$$I_\infty(X,Y) = \log \mathbb{E}_{P_Y}\left[\operatorname*{ess-sup}_{P_X} \frac{dP_{XY}}{dP_X P_Y}\right] = \mathcal{L}\left(X \to Y\right), \tag{2}$$

where $\mathcal{L}(X\!\to\!Y)$ denotes the maximal leakage from X to Y, a recently defined information measure with an operational meaning in the context of privacy and security (Issa et al., 2020). Maximal leakage represents the main quantity of interest for the scope of this paper, as it is amenable to analysis and has been used to bound the generalization error (Esposito et al., 2021). As such, we will bound the maximal leakage between the input and output of generic noisy iterative algorithms.

To that end, we mention a few useful properties of $\mathcal{L}(X\!\to\!Y)$. If X and Y are jointly continuous random variables, then (Issa et al., 2020, Corollary 4)

$$\mathcal{L}(X\!\to\!Y) = \log \int \underset{P_X}{\text{ess-sup}}\, f_{Y|X}(y|x)dy, \tag{3}$$

where $f_{Y|X}$ is the conditional pdf of Y given X. Moreover, maximal leakage satisfies the following chain rule (the proof of which is given in Appendix A):

Lemma 1 *Given a triple of random variables (X, Y_1, Y_2), then*

$$\mathcal{L}(X\!\to\!Y_1, Y_2) \leq \mathcal{L}(X\!\to\!Y_1) + \mathcal{L}(X\!\to\!Y_2|Y_1), \tag{4}$$

where the conditional maximal leakage $\mathcal{L}(X\!\to\!Y_2|Y_1) = \text{ess-sup}_{P_{Y_1}} \mathcal{L}(X\!\to\!Y_2|Y_1 = y_1)$, where the latter term is interpreted as the maximal leakage from X to Y_2 with respect to the distribution $P_{XY_2|Y_1=y_1}$. Consequently, for random variables $(X, (Y_i)_{i=1}^n)$,

$$\mathcal{L}(X\!\to\!Y^n) \leq \sum_{i=1}^{n} \mathcal{L}\left(X\!\to\!Y_i|Y^{i-1}\right). \tag{5}$$

Moreover, one can relate $\mathcal{L}(X\!\to\!Y)$ to $I(X;Y)$ through I_α. Indeed, an important property of I_α is that it is non-decreasing in α, hence for every $\infty > \alpha > 1$:

$$I(X;Y) = I_1(X,Y) \leq I_\alpha(X,Y) \leq I_\infty(X,Y) = \mathcal{L}(X\!\to\!Y). \tag{6}$$

For more details on Sibson's α-MI we refer the reader to (Verdú, 2015), as for maximal leakage the reader is referred to (Issa et al., 2020).

2.1.2. Learning Setting

Let $\mathcal{Z}$ be the sample space, $\mathcal{W}$ be the hypothesis space, and $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$ be a loss function. Say $\mathcal{W} \subseteq \mathbb{R}^d$. Let $S = (Z_1, Z_2, \ldots, Z_n)$ consist of n i.i.d samples, where $Z_i \sim P$, with P unknown. A learning algorithm $\mathcal{A}$ is a mapping $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ that given a sample S provides a hypothesis $W = \mathcal{A}(S)$. $\mathcal{A}$ can be either a deterministic or a randomized mapping and undertaking a probabilistic (and information-theoretic) approach one can then equivalently consider $\mathcal{A}$ as a family of conditional probability distributions $P_{W|S=s}$ for $s \in \mathcal{Z}^n$ *i.e.*, an information channel. Given a hypothesis $w \in \mathcal{W}$ the true risk of w is denoted as follows:

$$L_{P_Z}(w) = \mathbb{E}_P[\ell(w, Z)] \tag{7}$$

while the empirical risk of w on S is denoted as follows:

$$L_S(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i). \tag{8}$$

Given a learning algorithm $\mathcal{A}$, one can then define its generalization error as follows:

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) = L_{\mathcal{P}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S)). \tag{9}$$

Since both S and $\mathcal{A}$ can be random, $\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S)$ is a random variable and one can then study its expected value or its behavior in probability. Bounds on the expected value of the generalization error in terms of information measures are given in Xu and Raginsky (2017); Issa et al. (2019); Bu et al. (2019); Steinke and Zakynthinou (2020) stating different variants of the following bound (Xu and Raginsky, 2017, Theorem 1): if $\ell(w, Z)$ is σ^2-sub-Gaussian[1] then

$$|\mathbb{E}[\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S)]| \leq \sqrt{\frac{2\sigma^2 I(S; \mathcal{A}(S))}{n}}. \tag{10}$$

Thus, if one can prove that the mutual information between the input and output of a learning algorithm $\mathcal{A}$ trained on S is bounded (ideally, growing less than linearly in n) then the expected generalization error of $\mathcal{A}$ will vanish with the number of samples. Alternatively, Esposito et al. (2021) demonstrate high-probability bounds, involving different families of information measures. One such bound, which is relevant to the scope of this paper is the following (Esposito et al., 2021, Corollary 2): assume $\ell(w, Z)$ is σ^2-sub-Gaussian and let $\alpha > 1$, then

$$\mathbf{Pr}(|\text{gen-err}_P(\mathcal{A}, S)| \geq t) \leq 2\exp\left(-\frac{\alpha-1}{\alpha}\left(\frac{nt^2}{2\sigma^2} - I_\alpha(S, \mathcal{A}(S))\right)\right), \tag{11}$$

taking the limit of $\alpha \to \infty$ in (11) leads to the following (Esposito et al., 2021, Corollary 4):

$$\mathbf{Pr}(|\text{gen-err}_P(\mathcal{A}, S)| \geq t) \leq 2\exp\left(-\left(\frac{nt^2}{2\sigma^2} - \mathcal{L}\left(S \to \mathcal{A}(S)\right)\right)\right). \tag{12}$$

Thus, in this case, if one can prove that the maximal leakage between the input and output of a learning algorithm $\mathcal{A}$ trained on S is bounded, then the **probability** of the generalization error of $\mathcal{A}$ being larger than any constant t will decay **exponentially fast** in the number of samples n.

2.2. Problem Setup

We consider iterative algorithms, where each update is of the following form:

$$W_t = g(W_{t-1}) - \eta_t F(W_{t-1}, Z_t) + \xi_t, \; \forall\, t \geq 1, \tag{13}$$

where $Z_t \subseteq S$ (sampled according to some distribution), $g : \mathbb{R}^d \to \mathbb{R}^d$ is a deterministic function, $F(W_{t-1}, Z_t)$ computes a direction (e.g., gradient), η_t is the step-size, and $\xi_t = (\xi_{t1}, \ldots, \xi_{td})$ is random noise. We will assume for the remainder of this paper that ξ_t has an absolutely continuous distribution. Let T denote the total number of iterations, $W^t = (W_1, W_2, \ldots W_t)$, and $Z^t = (Z_1, Z_2, \ldots, Z_t)$. The algorithms under consideration further satisfy the following two assumptions

- **Assumption 1 (Sampling):** The sampling strategy is agnostic to parameter vectors:

$$P(Z_{t+1}|Z^t, W^t, S) = P(Z_{t+1}|Z^t, S). \tag{14}$$

[1] A 0-mean random variable X is said to be σ^2-sub-Gaussian if $\log \mathbb{E}[\exp(\lambda X)] \leq \sigma^2\lambda^2/2$ for every $\lambda \in \mathbb{R}$.

- **Assumption 2 ($\mathbf{L_p}$-Boundedness):** For some $p > 0$ and $L > 0$, $\sup_{w,z} \|F(w,z)\|_p \leq L$.

As a consequence of the first assumption and the structure of the iterates, we get:

$$P(W_{t+1}|W^t, Z^T, S) = P(W_{t+1}|W_t, Z_{t+1}). \tag{15}$$

The above setup was proposed by Pensia et al. (2018), who specifically studied the case $p = 2$. Denoting by W the final output of the algorithm (some function of W^T), they show that

Theorem 2 ((Pensia et al., 2018, Theorem 1)) *If the boundedness assumption holds for $p = 2$ and $\xi_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$, then*

$$I(S; W) \leq \frac{d}{2} \sum_{t=1}^{T} \log\left(1 + \frac{\eta_t^2 L^2}{d\sigma_t^2}\right). \tag{16}$$

By virtue of inequality (10), this yields a bound on the expected generalization error.

In this work, we derive bounds on the maximal leakage between $\mathcal{L}\left(S \to W\right)$ for iterative noisy algorithms, which leads to high-probability bounds on the generalization error (cf. equation (12)). We consider different scenarios in which F is bounded in L_1, L_2, or L_∞ norm, and the added noise is Laplace, Gaussian, or Uniform. It is worth noting that the bounds we derive depend on F only through the boundedness assumption (Assumption 2 above). Considering F to be a gradient yields the most (practically) interesting scenario in which our results hold, as it represents a widely used family of learning algorithms. However, we do not leverage any structure that is particular to gradients (beyond the boundedness assumption).

2.3. Notation

Given $d \in \mathbb{N}$, $w \in \mathbb{R}^d$, and $r > 0$, let $\mathcal{B}_p^d(w, r) = \{x \in \mathbb{R}^d : \|x - w\|_p \leq r\}$ denote the L_p-ball of radius r and center w, and let $V_p(d, r)$ denote its corresponding volume. When the dimension d is clear from the context, we may drop the superscript and write $\mathcal{B}_p(w, r)$. Given a set S, we denote its complement by $\overline{S}$. The i-th component of w_t will be denoted by w_{ti}.

We denote the pdf of the noise ξ_t by $f_t : \mathbb{R}^d \to \mathbb{R}$. The following functional will be useful for our study: given $d \in \mathbb{N}$, $p > 0$, a pdf $f : \mathbb{R}^d \to \mathbb{R}$, and an $r \geq 0$, define

$$h(d, p, f, r) := \int\limits_{\overline{\mathcal{B}_p^d}(0,r)} \sup_{x \in \mathcal{B}_p^d(0,r)} f(w - x) \mathrm{d}w. \tag{17}$$

We denote the "positive octant" by A_d, i.e.,

$$A_d := \{w \in \mathbb{R}^d : w_i \geq 0, \text{ for all } i \in \{1, 2, \ldots, d\}\}. \tag{18}$$

Since we will mainly consider pdfs that are symmetric (Gaussian, Laplace, uniform), the h functional "restricted" to A_d will be useful:

$$h_+(d, p, f, r) := \int\limits_{\overline{\mathcal{B}_p^d}(0,r) \cap A_d} \sup_{x \in \mathcal{B}_p^d(0,r)} f(w - x) \mathrm{d}w. \tag{19}$$

2.4. General Bound

Proposition 3 *Suppose $f_t : \mathbb{R}^d \to \mathbb{R}$ is maximized for $x = 0$. If Assumptions 1 and 2 hold for some $p > 0$, then*

$$\mathcal{L}\left(S{\to}W\right) \le \sum_{t=1}^{T} \log\left(f_t(0)V_p(d,\eta_t L) + h(d,p,f_t,\eta_t L)\right), \tag{20}$$

where h is defined in equation (17).

The above bound is appealing as it implicitly poses an optimization problem: given a constraint on the noise pdf f_t (say, a bounded variance), one may choose f_t as to minimize the upper bound in equation (20). Moreover, despite its generality, we show that it is tight in several interesting cases, including when $p = 2$ and f_t is the Gaussian pdf.

In the next section, we consider several scenarios for different values of p and different noise distributions. As a testament to the tractability of maximal leakage, we derive exact semi-closed form expressions for the bound of Proposition 3. Finally, it is worth noting that the form of the bound allows us to choose different noise distributions at different time steps, but these examples are outside the scope of this paper.

Proof We proceed as in the work of Pensia et al. (2018):

$$\mathcal{L}\left(S{\to}W\right) \le \mathcal{L}\left(Z^T{\to}W^T\right) \le \sum_{t=1}^{T}\mathcal{L}\left(Z^T{\to}W_t|W^{t-1}\right) = \sum_{t=1}^{T}\mathcal{L}\left(Z_t{\to}W_t|W_{t-1}\right), \tag{21}$$

where the first inequality follows from Lemma 2 of Pensia et al. (2018) and the data processing inequality for maximal leakage (Issa et al., 2020, Lemma 1), the second inequality follows Lemma 1, and the equality follows from (15). Now,

$$\exp\left\{\mathcal{L}\left(Z_t{\to}W_t|W_{t-1}=w_{t-1}\right)\right\} = \int_{\mathbb{R}^d} \operatorname*{ess-sup}_{P_{z_t}} p(w_t|Z_t)\mathrm{d}w_t \tag{22}$$

$$= \int_{\mathbb{R}^d} \operatorname*{ess-sup}_{P_{z_t}} f_t\left(w_t - g(w_{t-1}) + \eta_t F(w_{t-1},Z_t)\right)\mathrm{d}w_t, \tag{23}$$

$$= \int_{\mathbb{R}^d} \operatorname*{ess-sup}_{P_{z_t}} f_t\left(w_t + \eta_t F(w_{t-1},Z_t)\right)\mathrm{d}w_t, \tag{24}$$

where the last equality follows from a change of a variable $w_t \leftarrow w_t - g(w_{t-1})$. Finally, since $\eta_t F(w_{t-1},z_t) \in \mathcal{B}_p(0,\eta_t L)$ by assumption, we can further upper-bound the above by:

$$\exp\left\{\mathcal{L}\left(Z_t{\to}W_t|W_{t-1}=w_{t-1}\right)\right\} \tag{25}$$

$$\le \int_{\mathbb{R}^d} \sup_{x_t\in\mathcal{B}_p(0,\eta_t L)} f_t\left(w_t + x_t\right)\mathrm{d}w_t \tag{26}$$

$$= \int_{\mathcal{B}_p(0,\eta_t L)} \sup_{x_t\in\mathcal{B}_p(0,\eta_t L)} f_t\left(w_t + x_t\right)\mathrm{d}w_t + \int_{\overline{\mathcal{B}_p}(0,\eta_t L)} \sup_{x_t\in\mathcal{B}_p(0,\eta_t L)} f_t\left(w_t + x_t\right)\mathrm{d}w_t \tag{27}$$

$$= f_t(0)V_p(d,\eta_t L) + \int_{\overline{\mathcal{B}_p}(0,\eta_t L)} \sup_{x_t\in\mathcal{B}_p(0,\eta_t L)} f_t\left(w_t - x_t\right)\mathrm{d}w_t, \tag{28}$$

where the last equality follows from the assumptions on f_t. ■

3. Boundedness in L_2-Norm

Considering the case where F computes a gradient, then boundedness in L_2-norm is a common assumption. It is commonly enforced, for instance, using gradient clipping (Abadi et al., 2016a,b; Chen et al., 2020).

The case in which $p = 2$ and the noise is Gaussian leads to the strongest result in this paper:

Theorem 4 *If the boundedness assumption holds for $p \leq 2$ and $\xi_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$, then*

$$\mathcal{L}(S \to W) \leq \sum_{t=1}^{T} \log \left(\frac{V_2(d, \eta_t L)}{(2\pi\sigma_t^2)^{d/2}} + \frac{1}{\Gamma\left(\frac{d}{2}\right)} \sum_{i=0}^{d-1} \Gamma\left(\frac{i+1}{2}\right) \left(\frac{\eta_t L}{\sigma_t \sqrt{2}}\right)^{d-1-i} \right), \tag{29}$$

where $V_2(d, r) = \dfrac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} r^d$. Consequently, for fixed T,

$$\lim_{d\to\infty} \mathcal{L}(S \to W) = 0. \tag{30}$$

Remarkably, equation (30) states that, as d grows, S and W are asymptotically independent. The bound is asymptotically optimal for $\mathcal{L}(S \to W)$ (indeed it yields an equality). More importantly, the induced high probability bound by equation (12) is also optimal. Indeed, at the limit when S and W are independent, the bound (12) recovers (the order optimal) McDiarmid's inequality *i.e.*, under the assumptions of Theorem 4 and considering the $0-1$ loss:

$$\mathbf{Pr}(|\text{gen-err}(\mathcal{A}, S)| \geq t) \leq 2\exp(-2nt^2). \tag{31}$$

This can be seen as an explanation of the (arguably un-intuitive) phenomenon that deeper networks often generalize better (also analyzed by Wang et al. (2023)).

By contrast, this is not captured in the bound by Pensia et al. (2018) given in equation (16). Indeed, it is growing as a function of d, and tends to a non-zero value:

$$\lim_{d\to\infty} \frac{d}{2} \sum_{t=1}^{T} \log\left(1 + \frac{\eta_t^2 L^2}{d\sigma_t^2}\right) = \sum_{t=1}^{T} \frac{\eta_t^2 L^2}{2\sigma_t^2}, \tag{32}$$

where the equality follows from the fact that $\lim_{n\to\infty} (1 + c/n)^c = e^c$ for all $c \in \mathbb{R}$. Notably, however, $I(S; W) \leq \mathcal{L}(S \to W)$ (cf. equation (6)), so that for large d, we have

$$I(S; W) \leq \mathcal{L}(S \to W) \leq \text{right-hand side of (29)} \leq \text{ right-hand side of (16)}. \tag{33}$$

As such, the bound in Theorem 4 is also a tighter upper bound on $I(S; W)$ for large enough d.

Moreover, note that even if the parameter L is large (e.g., Lipschitz constant of a neural network (Negrea et al., 2019)), it appears in (29) normalized by $\Gamma(d/2)$ so its effect is significantly dampened (as d is also typically very large).

Finally, note that the bound in Proposition 3 is increasing in p: this can be seen from line (26), where the supremum over $\mathcal{B}_p$ can be further upper-bounded by a supremum over $\mathcal{B}_{p'}$ for $p' > p$.

Therefore for $q \leq p$, the bound induced by Proposition 3 is smaller. The bound in Theorem 4 corresponds to $p = 2$ and goes to 0 (as d grows), hence the bound induced by Proposition 3 goes to 0 for all $q \leq p = 2$. In particular, adding Gaussian noise is asymptotically optimal (in the sense discussed above) when Assumption 2 holds for any $p \leq 2$.

Proof To show that the right hand side of Equation (29) goes to zero as $d \to \infty$, we use Stirling's approximation of the Gamma function: for all $x > 0$,

$$\sqrt{2\pi} x^{x-\frac{1}{2}} e^{-x} \leq \Gamma(x) \leq \sqrt{2\pi} x^{x-\frac{1}{2}} e^{-x} e^{\frac{1}{12x}}. \tag{34}$$

The details of the computation can be found in Appendix B. We now turn to the proof of inequality (29). The conditions of Proposition 3 are satisfied, thus it is sufficient to prove the bound for $p = 2$ (cf. discussion above):

$$\mathcal{L}\left(S \to W\right) \leq \sum_{t=1}^{T} \log \left(f_t(0) V_2(d, \eta_t L) + \int\limits_{\overline{\mathcal{B}_2}(0,\eta_t L)} \sup_{x_t \in \mathcal{B}_2(0,\eta_t L)} f_t(w_t - x_t) \mathrm{d}w_t \right) \tag{35}$$

$$= \sum_{t=1}^{T} \log \left(\frac{V_2(d, \eta_t L)}{(2\pi\sigma_t^2)^{\frac{d}{2}}} + \int\limits_{\overline{\mathcal{B}_2}(0,\eta_t L)} \sup_{x_t \in \mathcal{B}_2(0,\eta_t L)} \frac{1}{(2\pi\sigma_t^2)^{\frac{d}{2}}} \exp\left\{ -\frac{\|w_t - x_t\|_2^2}{2\sigma_t^2} \right\} \mathrm{d}w_t \right). \tag{36}$$

Hence, it remains to show that the second term inside the $\log$ matches that of equation (29). To that end, note that the point in $\mathcal{B}_2(0, \eta_t L)$ that minimizes the distance to w_t is given $\frac{\eta_t L}{\|w_t\|} w_t$. So we get

$$\|w_t - x_t\| \geq \|w_t - \frac{\eta_t L}{\|w_t\|} w_t\| = \|w_t\| - \eta_t L. \tag{37}$$

Then,

$$h(d, 2, f_t, \eta_t L) = \int\limits_{\overline{\mathcal{B}_2}(0,\eta_t L)} \sup_{x_t \in \mathcal{B}_2(0,\eta_t L)} \frac{1}{(2\pi\sigma_t^2)^{\frac{d}{2}}} \exp\left\{ -\frac{\|w_t - x_t\|_2^2}{2\sigma_t^2} \right\} \mathrm{d}w_t \tag{38}$$

$$= \int\limits_{\overline{\mathcal{B}_2}(0,\eta_t L)} \frac{1}{(2\pi\sigma_t^2)^{\frac{d}{2}}} \exp\left\{ -\frac{(\|w_t\|_2 - \eta_t L)^2}{2\sigma_t^2} \right\} \mathrm{d}w_t. \tag{39}$$

To evaluate this integral, we use spherical coordinates (details in Appendix C). Then,

$$h(d, 2, f_t, \eta_t L) = \left(\frac{\eta_t L}{\sigma_t \sqrt{2}} \right)^{d-1} \frac{1}{\Gamma\left(\frac{d}{2}\right)} \sum_{i=0}^{d-1} \left(\frac{\sigma_t \sqrt{2}}{\eta_t L} \right)^{i} \Gamma\left(\frac{i+1}{2} \right). \tag{40}$$

Combining equations (36) and (40) yields (29). ■

Remark 5 *One could also derive a semi-closed form bound for the case in which the added noise is uniform. However, in that case $\mathcal{L}\left(S \to W\right)$ goes to infinity as d goes to infinity. The same behavior holds if the added noise is Laplace. Since the Gaussian noise leads to an asymptotically optimal bound, we skip the analysis of uniform and Laplace noise.*

4. Boundedness in L_∞-Norm

The bound in Proposition 3 makes minimal assumptions about the pdf f_t. In many practical scenarios we have more structure we could leverage. In particular, we make the following standard assumptions in this section:

- ξ_t is composed of i.i.d components. Let f_{t0} be the pdf of a component, then $f_t(x_t) = \prod_{i=1}^{d} f_{t0}(x_{ti})$.
- f_{t0} is symmetric around 0 and non-increasing over $[0, \infty)$.

In this setting, Proposition 3 reduces to a very simple form for $p = \infty$:

Theorem 6 *Suppose f_t satisfies the above assumptions. If Assumptions 1 and 2 hold for $p = \infty$, then*

$$\mathcal{L}\left(S \to W\right) \leq \sum_{t=1}^{T} d \log\left(1 + 2\eta_t L f_{t0}(0)\right). \tag{41}$$

Unlike the bound of Theorem 4, the limit as d goes to infinity here is infinite. However, the bounded-L_∞ assumption is *weaker* than the bounded L_2-norm assumption. Moreover, the assumption of having a bounded L_∞-norm is satisfied in Pichapati et al. (2019) where the authors clipped the gradient in terms of the L_∞-norm, thus "enforcing" the assumption. On the other hand, the theorem has an intriguing form as, under standard assumptions, the bound depends on f_{t0} only through $f_{t0}(0)$. This naturally leads to an optimization problem: given a certain constraint on the noise, which distribution $f^\star$ minimizes $f(0)$? The following theorem shows that, if the noise is required to have a bounded variance, then $f^\star$ corresponds to the uniform distribution:

Theorem 7 *Let $\mathcal{F}$ be the family of probability densities (over $\mathbb{R}$) satisfying for each $f \in \mathcal{F}$:*

1. *f is symmetric around 0.*
2. *f is non-increasing over $[0, \infty)$.*
3. *$\mathbf{E}_f[X^2] \leq \sigma^2$.*

Then, the distribution minimizing $f(0)$ over $\mathcal{F}$ is the uniform distribution $\mathcal{U}(-\sigma\sqrt{3}, \sigma\sqrt{3})$.

That is, uniform noise is optimal in the sense that it minimizes the upper bound in Theorem 6 under bounded variance constraints. The proof of Theorem 7 is deferred to Appendix E.

4.1. Proof of Theorem 6

Since the assumptions of Proposition 3 hold, then

$$\mathcal{L}\left(S \to W\right) \leq \sum_{t=1}^{T} \log\left(f_t(0) V_\infty(d, \eta_t L) + \int\limits_{\overline{\mathcal{B}_\infty}(0,\eta_t L)} \sup_{x_t \in \mathcal{B}_\infty(0,\eta_t L)} f_t(w_t - x_t) \mathrm{d}w_t \right) \tag{42}$$

$$= \sum_{t=1}^{T} \log\left((2\eta_t L f_{t0}(0))^d + \int\limits_{\overline{\mathcal{B}_\infty}(0,\eta_t L)} \prod_{i=1}^{d} \sup_{x_{ti}:|x_{ti}| \leq \eta_t L} f_{t0}(w_{ti} - x_{ti}) \mathrm{d}w_t \right). \tag{43}$$

It remains to show that $h(d, \infty, f_t, \eta_t L)$ (i.e., the second term inside the $\log$ in Equation (17)) is equal to $(1 + 2\eta_t L f_{t0}(0))^d - (2\eta_t L f_{t0}(0))^d$. We will derive a recurrence relation for h in terms of d. To simplify the notation, we drop the subscript t and ignore the dependence of h on $p = \infty$, f_t, and $\eta_t L$, so that we simply write $h(d)$ (and correspondingly, $h_+(d)$, cf. Equation (19)).

By symmetry, $h(d) = 2^d h_+(d)$. Letting $w^{d-1} := (w_1, \ldots, w_{d-1})$, we will decompose the integral over $\overline{\mathcal{B}^d_\infty}(0, \eta_t L)$ into two disjoint subsets: 1) $w^{d-1} \notin \mathcal{B}^{d-1}_\infty(0, \eta_t L)$, in which case w_d can take any value in $\mathbb{R}$, and 2) $w^{d-1} \in \mathcal{B}^{d-1}_\infty(0, \eta_t L)$, in which case w_d must satisfy $|w_d| > \eta_t L$.

$$h_+(d) = \int\limits_{\overline{\mathcal{B}^{d-1}_\infty}(0,\eta_t L) \cap A_{d-1}} \prod_{i=1}^{d-1} \sup_{x_i:|x_i| \leq \eta_t L} f(w_i - x_i) \int_0^\infty \sup_{x_d:|x_d| \leq \eta_t L} f(w_d - x_d) \mathrm{d}w_d \mathrm{d}w^{d-1} \tag{44}$$

$$+ \int\limits_{\mathcal{B}^{d-1}_\infty(0,\eta_t L) \cap A_{d-1}} \prod_{i=1}^{d-1} \sup_{x_i:|x_i| \leq \eta_t L} f(w_i - x_i) \int_{\eta_t L}^\infty \sup_{x_d:|x_d| \leq \eta_t L} f(w_d - x_d) \mathrm{d}w_d \mathrm{d}w^{d-1} \tag{45}$$

The innermost integral of line (45) is independent of w^{d-1} so that the outer integral is equal to $h_+(d-1)$. Similarly, the innermost integral of line (44) is independent of w^{d-1}, and the supremum in the outer integral yields $f(0)$ for every i. Hence, we get

$$h(d) = (1 + 2\eta_t L f(0))\, h(d-1) + (2\eta_t L f(0))^{d-1}, \tag{46}$$

the detailed proof of which is deferred to Appendix D. Finally, it is straightforward to check that $h(1) = 1$, hence $h(d) = (1 + 2\eta_t L f(0))^d - (2\eta_t L f(0))^d$.

5. Boundedness in L_1-Norm

In this section, we consider the setting where Assumption 2 holds for $p = 1$. By Proposition 3, any bound derived for $p = 2$ holds for $p = 1$ as well. In particular, Theorem 4 applies so that $\mathcal{L}\left(S \to W\right)$ goes to zero when the noise is Gaussian. Nevertheless, it is possible to compute a semi-closed form directly for $p = 1$ (cf. Theorem 9 below).

Considering the optimality of Gaussian noise for the $p = 2$ case, and the optimality of uniform noise (in the sense discussed above) for $p = \infty$ case, one might wonder if Laplace noise also minimizes the bound of Proposition 3 for the $p = 1$ case. We answer this question in the negative, as the limit of the leakage in this case is a non-zero constant (cf. Theorem 8), as opposed to the zero limit when the noise is Gaussian.

5.1. Bound for Laplace noise

We say X has a Laplace distribution, denoted by $X \sim \text{Lap}(\mu, 1/\lambda)$, if its pdf is given by $f(x) = \frac{\lambda}{2} e^{-\lambda|x-\mu|}$ for $x \in \mathbb{R}$, for some $\mu \in \mathbb{R}$ and $\lambda > 0$. The corresponding variance is given by $2/\lambda^2$.

Theorem 8 *If the boundedness assumptions holds for $p = 1$ and ξ_t is composed of i.i.d components, each of which is $\sim \text{Lap}(0, \frac{\sigma_t}{\sqrt{2}})$, then*

$$\mathcal{L}\left(S \to W\right) \leq \sum_{t=1}^{T} \log\left(\frac{V_1(d, \eta_t L)}{(\sigma_t\sqrt{2})^d} + \sum_{i=0}^{d-1} \frac{(\sigma_t \eta_t L/\sqrt{2})^i}{i!}\right), \tag{47}$$

where $V_1(d, r) = \dfrac{(2r)^d}{d!}$. Consequently, for fixed T,

$$\lim_{d\to\infty} \mathcal{L}\left(S \to W\right) \leq \sum_{t=1}^{T} \frac{\sigma_t \eta_t L}{\sqrt{2}}. \tag{48}$$

Proof We give a high-level description of the proof (as similar techniques have been used in proofs of earlier theorems) and defer the details to Appendix F. Since the multivariate Laplace distribution (for i.i.d variables) depends on the L_1-norm of the corresponding vector of variables, we need to solve the following problem: given $R > 0$ and $w \notin \mathcal{B}_1(0, R)$, compute

$$\inf_{x \in \mathcal{B}_1(0,R)} \|w - x\|_1. \tag{49}$$

The closest element in $\mathcal{B}_1(0, R)$ will lie on the hyperplane defining $\mathcal{B}_1$ that is in the same octant as w, so the problem reduces to projecting a point on a hyperplane in L_1-distance (the proof in the appendix does not follow this argument but arrives at the same conclusion). Then, we need to compute $h(d, 1, f_t, \eta_t L)$. We use a similar approach as in the proof of Theorem 6, that is, we split the integral and derive a recurrence relation. ■

5.2. Bound for Gaussian noise

Finally, we derive a bound on the induced leakage when the added noise is Gaussian:

Theorem 9 *If the boundedness assumptions holds for $p = 1$ and $\xi_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$, then*

$$\mathcal{L}\left(S \to W\right) \leq \sum_{t=1}^{T} \log\left(\frac{V_1(d, R_t)}{(2\pi\sigma^2)^{\frac{d}{2}}} + \frac{(2\eta_t L)^{d-1}(\sigma_t\sqrt{2d})}{(2\pi\sigma_t^2)^{\frac{d}{2}}((d-1)!)} \sum_{i=0}^{d-1} \left(\frac{\sigma_t\sqrt{2d}}{\eta_t L}\right)^i \Gamma\left(\frac{i+1}{2}\right)\right). \tag{50}$$

In order to prove Theorem 9 one has to solve a problem similar to the one introduced in Theorem 8 (cf. equation (49)). However, in this case a different norm is involved: i.e., given $R > 0$ and $w \notin \mathcal{B}_1(0, R)$, one has to compute

$$\inf_{x \in \mathcal{B}_1(0,R)} \|w - x\|_2. \tag{51}$$

Again, one can argue that the point achieving the infimum lies on the hyperplane defining $\mathcal{B}_1$ that is in the same octant as w. In other words, the minimizer $x^\star$ is such that the sign of each component is the same sign as the corresponding component of w (and lies on the boundary of $\mathcal{B}_1$). Thus, we are projecting a point on the corresponding face of the L_1-ball. The length of the projection is then appropriately lower-bounded and the induced integral is solved by an opportune choice of change of variables. The details of the proof are given in Appendix G.

6. Conclusions

In this work, we analyzed the Maximal Leakage of SGLD-like mechanisms. The motivation behind this analysis is the relationship between having a bounded leakage and exponential concentration of the generalization error of the learning algorithm (Esposito et al., 2021). Moreover, with additional assumptions over the loss function, one can leverage the ordering between mutual information and Sibson's α-Mutual Information to automatically provide bounds on the expected generalization error as well. Our initial contribution is the introduction of a general bound on maximal leakage (Proposition 3) which depends solely on the L_p-Boundedness assumption (*e.g.*, of the gradient of the loss) and the pdf of the noise that is added in the iterates of the algorithm (Equation (13)). As a consequence of such bound, we could explicitly upper bound maximal leakage in a variety of settings while shedding some light on the influence of the boundedness assumption on the performance of the algorithm. For instance, in Section 3 we proved that if one asks for the L_p-norm with $p \leq 2$ to be bounded, then adding Gaussian noise leads to an *asymptotically optimal*[2] behavior of the bound: the maximal leakage in this setting goes to 0 as the number of parameters d approaches infinity – that is, the input and output of the learning algorithm become statistically independent as the number of parameters grows, leading to an exponentially vanishing probability of generalization error. However, if one adds uniform or Laplacian noise, the same result does not seem to hold and the bound on maximal leakage goes to infinity as d grows. Similarly, we analyze the choice of noise to be added (among a family of allowed noise distributions, e.g., satisfying a variance constraint) when one asks for the L_∞-norm to be bounded (in order to minimize the *bound* on maximal leakage under bounded variance constraints). Hence, one can summarize the main contributions in the following way:

- Our analysis, other than proving that some SGLD-like algorithms generalize well, provides an *information-theoretic* interpretation of why over-parametrized models (with a number of parameters d much greater than the number of samples n) have better generalization guarantees;

- analyzing maximal leakage can also inform the design of novel iterative algorithms as it allows to explicitly link the assumptions and structure of the iterates to the bound one can retrieve on the generalization error. For instance, we saw how adding Gaussian noise is asymptotically optimal whenever one assumes the L_p-norm of the gradient to be bounded with $p \leq 2$ while adding Laplacian or uniform noise, in this case, may not lead to the same guarantees.

[2] **asymptotic** in the number of parameters d, **optimal** decay on the probability of having a large generalization error in the number of samples n

Acknowledgment

The work in this manuscript was supported in part by the Swiss National Science Foundation under Grant 200364 and by the University Research Board at the American University of Beirut (Beirut, Lebanon).

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016a.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016b.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482, mar 2003. ISSN 1532-4435.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/b22b257ad0519d4500539da3c8bcf4dd-Paper.pdf.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. URL http://jmlr.org/papers/v20/17-612.html.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. volume 83 of *Proceedings of Machine Learning Research*, pages 25–55. PMLR, 07–09 Apr 2018.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer, 2003. ISBN 3-540-23122-6.

Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 587–591, 2019. doi: 10.1109/ISIT.2019.8849590.

Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 121–130, 2020. doi: 10.1109/JSAIT.2020.2991139.

Xiangyi Chen, Steven Z. Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33:13773–13782, 2020.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *CoRR*, abs/1703.11008, 2017. URL https://arxiv.org/abs/1703.11008.

Amedeo Roberto Esposito and Michael Gastpar. From generalisation error to transportation-cost inequalities and back. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 294–299, 2022. doi: 10.1109/ISIT50566.2022.9834354.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8): 4986–5004, 2021. doi: 10.1109/TIT.2021.3085190.

Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16457–16467. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/befe5b0172188ad14d48c3ebe9cf76bf-Paper.pdf.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M. Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms, 2020. URL https://arxiv.org/abs/2004.12983.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839, 2020. doi: 10.1109/JSAIT.2020.3040992.

Ibrahim Issa, Amedeo Roberto Esposito, and Michael Gastpar. Strengthened information-theoretic bounds on the generalization error. In *2019 IEEE International Symposium on Information Theory, ISIT Paris, France, July 7-12*, 2019.

Ibrahim Issa, Aaron B. Wagner, and Sudeep Kamath. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657, 2020. doi: 10.1109/TIT.2019.2962804.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Gabor Lugosi and Gergely Neu. Generalization bounds via convex analysis. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3524–3546. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/lugosi22a.html.

David A. McAllester. A pac-bayesian tutorial with A dropout bound. *CoRR*, abs/1307.2118, 2013. URL http://arxiv.org/abs/1307.2118.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/05ae14d7ae387b93370d142d82220f1b-Paper.pdf.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3526–3545. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/neu21a.html.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550, 2018.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240. PMLR, 09–11 May 2016.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Robin Sibson. Information radius. *Z. Wahrscheinlichkeitstheorie verw Gebiete 14*, pages 149–160, 1969.

Thomas Steinke and Lydia Zakynthinou. Reasoning About Generalization via Conditional Mutual Information. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/steinke20a.html.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback-keibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, July 2014.

Vladimir N. Vapnik and Alexey Y. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis 1*, (3), 1991.

Sergio Verdú. α-mutual information. In *2015 Information Theory and Applications Workshop, ITA 2015, San Diego, CA, USA, February 1-6, 2015*, pages 1–6, 2015.

Bohan Wang, Huishuai Zhang, Jieyu Zhang, Qi Meng, Wei Chen, and Tie-Yan Liu. Optimizing information-theoretical generalization bound via anisotropic noise of SGLD. In

Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 26080–26090, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/db2b4182156b2f1f817860ac9f409ad7-Abstract.html.

Hao Wang, Rui Gao, and Flavio P Calmon. Generalization bounds for noisy iterative algorithms using properties of additive noise channels. *Journal of Machine Learning Research*, 24(26):1–43, 2023.

Aolin Xu and Maxin Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, page 2521–2530, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, feb 2021. ISSN 0001-0782. doi: 10.1145/3446776. URL https://doi.org/10.1145/3446776.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. In *International Conference on Learning Representations*, 2018.

Appendix A. Proof of Lemma 1

Recall the definition of maximal leakage and conditional maximal leakage:

Definition 10 (Maximal Leakage (Issa et al., 2020, Definition 1)) *Given two random variables (X, Y) with joint distribution P_{XY},*

$$\mathcal{L}\left(X \!\to\! Y\right) = \log \sup_{U: U - X - Y} \frac{\mathbf{Pr}(\hat{U}(Y) = U)}{\max_u P_U(u)}, \tag{52}$$

where U takes values in a finite, but arbitrary, alphabet, and $\hat{U}(Y)$ is the optimal estimator (i.e., MAP) of U given Y.

Similarly,

Definition 11 (Conditional Maximal Leakage (Issa et al., 2020, Definition 6)) *Given three random variables (X, Y, Z) with joint distribution P_{XYZ},*

$$\mathcal{L}\left(X \!\to\! Y|Z\right) = \log \sup_{U: U - X - Y|Z} \frac{\mathbf{Pr}(\hat{U}(Y, Z) = U)}{\mathbf{Pr}(\hat{U}(Z) = U)}, \tag{53}$$

where U takes values in a finite, but arbitrary, alphabet, and $\hat{U}(Y, Z)$ and $\hat{U}(Z)$ are the optimal estimators (i.e., MAP) of U given (Y, Z) and U given Z, respectively.

It then follows that

$$\begin{aligned}
\mathcal{L}\left(X \!\to\! Y_1, Y_2\right) &= \log \sup_{U: U - X - (Y_1, Y_2)} \frac{\mathbf{Pr}(\hat{U}(Y_1, Y_2) = U)}{\max_u P_U(u)} && (54)\\
&= \log \sup_{U: U - X - (Y_1, Y_2)} \frac{\mathbf{Pr}(\hat{U}(Y_1, Y_2) = U)}{\mathbf{Pr}(\hat{U}(Y_1) = U)} \frac{\mathbf{Pr}(\hat{U}(Y_1) = U)}{\max_u P_U(u)} && (55)\\
&\leq \log \sup_{U: U - X - (Y_1, Y_2)} \frac{\mathbf{Pr}(\hat{U}(Y_1, Y_2) = U)}{\mathbf{Pr}(\hat{U}(Y_1) = U)} \cdot \sup_{U: U - X - (Y_1, Y_2)} \frac{\mathbf{Pr}(\hat{U}(Y_1) = U)}{\max_u P_U(u)} && (56)\\
&\leq \log \sup_{U: U - X - Y_2|Y_1} \frac{\mathbf{Pr}(\hat{U}(Y_1, Y_2) = U)}{\mathbf{Pr}(\hat{U}(Y_1) = U)} \cdot \sup_{U: U - X - Y_1} \frac{\mathbf{Pr}(\hat{U}(Y_1) = U)}{\max_u P_U(u)} && (57)\\
&= \mathcal{L}\left(X \!\to\! Y_2|Y_1\right) + \mathcal{L}\left(X \!\to\! Y_1\right), && (58)
\end{aligned}$$

where the last inequality follows from the fact that $U - X - (Y_1, Y_2)$ implies $U - X - Y_2|Y_1$.

The fact that

$$\mathcal{L}\left(X \!\to\! Y_2|Y_1\right) = \underset{P_{Y_1}}{\text{ess-sup}}\, \mathcal{L}\left(X \!\to\! Y_2|Y_1 = y_1\right), \tag{59}$$

has been shown for discrete alphabets in Theorem 6 of (Issa et al., 2020). The extension to continuous alphabets is similar (with integrals replacing sums, and pdfs replacing pmfs, where appropriate).

Finally, it remains to show equation (5). We proceed by induction. The case $n = 2$ has already been shown above. Assume the inequality is true up to $n - 1$ variables, then

$$
\begin{aligned}
\mathcal{L}\left(X \to Y^n\right) &\leq \mathcal{L}\left(X \to Y_1\right) + \operatorname*{ess-sup}_{P_{Y_1}} \mathcal{L}\left(X \to Y_2^n | Y_1 = y_1\right) && (60) \\
&\leq \mathcal{L}\left(X \to Y_1\right) + \operatorname*{ess-sup}_{P_{Y_1}} \sum_{i=2}^{n} \mathcal{L}\left(X \to Y_i | Y^{i-1}, Y_1 = y_1\right) && (61) \\
&= \sum_{i=1}^{n} \mathcal{L}\left(X \to Y_i | Y^{i-1}\right), && (62)
\end{aligned}
$$

where the second inequality follows from the induction hypothesis.

Appendix B. Proof of equation (30)

For notational convenience, let $c_1 = \frac{\sigma_t \sqrt{2}}{\eta_t L}$ and $c_2 = \frac{2e}{c_1^2}$. Then,

$$
\begin{aligned}
\sum_{i=0}^{d-1} \frac{\Gamma\left(\frac{i+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} c_1^{i-(d-1)} &= 1 + \sum_{i=0}^{d-2} \frac{\Gamma\left(\frac{i+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \leq 1 + \sum_{i=0}^{d-2} c_1^{(i-(d-1))} \frac{\left(\frac{i+1}{2}\right)^{\frac{i}{2}} e^{-\frac{i+1}{2}} e^{\frac{1}{12i}}}{\left(\frac{d}{2}\right)^{\frac{d-1}{2}} e^{-\frac{d}{2}}} && (63) \\
&= 1 + e^{\frac{1}{12}} \left(\frac{2e}{c_1^2 d}\right)^{\frac{d-1}{2}} \sum_{i=0}^{d-2} \left(\frac{(i+1)c_1^2}{2e}\right)^{\frac{i}{2}} && (64) \\
&\leq 1 + e^{\frac{1}{12}} \left(\frac{c_2}{d}\right)^{\frac{d-1}{2}} \sum_{i=0}^{d-2} \left(\frac{d}{c_2}\right)^{\frac{i}{2}} && (65) \\
&= 1 + e^{\frac{1}{12}} \left(\frac{c_2}{d}\right)^{\frac{d-1}{2}} \frac{\left(\frac{d}{c_2}\right)^{\frac{d-1}{2}} - 1}{\sqrt{\frac{d}{c_2}} - 1} && (66) \\
&= 1 + e^{\frac{1}{12}} \frac{1 - \left(\frac{c_2}{d}\right)^{\frac{d-1}{2}}}{\sqrt{\frac{d}{c_2}} - 1} \xrightarrow{d \to \infty} 1. && (67)
\end{aligned}
$$

Moreover,

$$
\frac{V_2(d, \eta_t L)}{(2\pi\sigma_t^2)^{d/2}} = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2} + 1\right)} \left(\frac{\eta_t L}{\sqrt{2\pi\sigma_t^2}}\right)^d = V_2\left(d, \frac{\eta_t L}{\sqrt{2\pi\sigma_t^2}}\right) \xrightarrow{d \to \infty} 0. \qquad (68)
$$

Combining equations (67) and (68) yields the desired limit.

Appendix C. Proof of equation (40)

To evaluate the integral in line (39), we write it in spherical coordinates:

$$\begin{aligned}
&h(d,2,f_t,\eta_t L)\\
&= \int\limits_{\overline{\mathcal{B}_2}(0,\eta_t L)} \frac{1}{(2\pi\sigma_t^2)^{\frac{d}{2}}} \exp\left\{-\frac{(\|w_t\|_2-\eta_t L)^2}{2\sigma_t^2}\right\} \mathrm{d}w_t.\\
&= \frac{1}{(2\pi\sigma_t^2)^{\frac{d}{2}}} \int_0^{2\pi}\int_0^{\pi}\dots\int_0^{\pi}\int_{\eta_t L}^{\infty} e^{\frac{-(\rho-\eta_t L)^2}{2\sigma_t^2}} \rho^{d-1}\sin^{d-2}(\phi_1)\sin^{d-3}(\phi_2)\dots\sin(\phi_{d-2})\mathrm{d}\rho\mathrm{d}\phi_1^{d-1}\\
&= \frac{2\pi}{(2\pi\sigma_t^2)^{\frac{d}{2}}} \left(\int_0^{\pi}\sin^{d-2}(\phi_1)\mathrm{d}\phi_1\right)\dots\left(\int_0^{\pi}\sin(\phi_{d-2})\mathrm{d}\phi_{d-2}\right)\left(\int_{\eta_t L}^{\infty} e^{\frac{-(\rho-\eta_t L)^2}{2\sigma_t^2}}\rho^{d-1}\mathrm{d}\rho\right). \qquad (69)
\end{aligned}$$

Now, note that for any $n \in \mathbb{N}$, $\int_0^{\pi}\sin^n(x)\mathrm{d}x = 2\int_0^{\pi/2}\sin^n(x)\mathrm{d}x$, and

$$\begin{aligned}
\int_0^{\pi/2}\sin^n(x)\mathrm{d}x &\overset{\text{(a)}}{=} \int_0^1 \frac{u^n}{\sqrt{1-u^2}}\mathrm{d}u \overset{\text{(b)}}{=} \frac{1}{2}\int_0^1 t^{\frac{n-1}{2}}(1-t)^{-\frac{1}{2}}\mathrm{d}y \overset{\text{(c)}}{=} \frac{1}{2}\mathrm{Beta}\left(\frac{n+1}{2},\frac{1}{2}\right)\\
&= \frac{\sqrt{\pi}\Gamma\left(\frac{n+1}{2}\right)}{2\Gamma\left(\frac{n}{2}+1\right)}, \qquad (70)
\end{aligned}$$

where (a) follows from the change of variable $u = \sin x$, (b) follows from the change of variable $t = u^2$, (c) follows from the definition of the Beta function: $\mathrm{Beta}(s_1,s_2) = \int_0^1 t^{s_1-1}(1-t)^{s_2-1}$, and the last equality is a known property of the Beta function ($\Gamma(1/2)=\sqrt{\pi}$). Consequently,

$$\begin{aligned}
&2\pi\left(\int_0^{\pi}\sin^{d-2}(\phi_1)\mathrm{d}\phi_1\right)\dots\left(\int_0^{\pi}\sin(\phi_{d-2})\mathrm{d}\phi_{d-2}\right)\\
&= (2\pi)\prod_{i=1}^{d-2}\frac{\sqrt{\pi}\Gamma\left(\frac{i+1}{2}\right)}{\Gamma\left(\frac{i}{2}+1\right)} = (2\pi)\pi^{\frac{d-2}{2}}\frac{\Gamma(1)}{\Gamma(d/2)} = 2\pi^{d/2}\frac{1}{\Gamma(d/2)}. \qquad (71)
\end{aligned}$$

To evaluate the innermost integral, the following identity will be useful:

$$\int_0^{\infty} x^n e^{-x^2}dx = \frac{1}{2}\int_0^{\infty} t^{\frac{n+1}{2}}e^{-t}dt = \frac{\Gamma\left(\frac{n+1}{2}\right)}{2}, \qquad (72)$$

where the first equality follows from the change of variable $t = x^2$. Then,

$$\int_{\eta_t L}^{\infty} e^{\frac{-(\rho-\eta_t L)^2}{2\sigma_t^2}} \rho^{d-1} d\rho = \int_0^{\infty} e^{\frac{-\rho^2}{2\sigma_t^2}} (\rho + \eta_t L)^{d-1} d\rho \tag{73}$$

$$= \int_0^{\infty} \sum_{i=0}^{d-1} \binom{d-1}{i} (\eta_t L)^{d-1-i} \rho^i e^{\frac{-\rho^2}{2\sigma_t^2}} d\rho \tag{74}$$

$$\overset{\text{(a)}}{=} \sum_{i=0}^{d-1} \binom{d-1}{i} (\eta_t L)^{d-1-i} \int_0^{\infty} \left(\sigma_t \sqrt{2}\right)^{i+1} t^i e^{-t^2} d\rho \tag{75}$$

$$\overset{\text{(b)}}{=} (\eta_t L)^{d-1} (\sigma_t \sqrt{2}) \sum_{i=0}^{d-1} \left(\frac{\sigma_t \sqrt{2}}{\eta_t L} \right)^i \frac{\Gamma((i+1)/2)}{2}. \tag{76}$$

where (a) follows from the change of variable $t = \rho/(\sigma\sqrt{2})$, and (b) follows from (72).

Finally, combining equations (69), (71), and (76), we get

$$h(d, 2, f_t, \eta_t L) = \frac{2\pi^{d/2}}{(2\pi\sigma_t^2)^{\frac{d}{2}} \Gamma(d/2)} (\eta_t L)^{d-1} (\sigma_t \sqrt{2}) \sum_{i=0}^{d-1} \left(\frac{\sigma_t \sqrt{2}}{\eta_t L} \right)^i \frac{\Gamma((i+1)/2)}{2} \tag{77}$$

$$= \left(\frac{\eta_t L}{\sigma_t \sqrt{2}} \right)^{d-1} \frac{1}{\Gamma(d/2)} \sum_{i=0}^{d-1} \left(\frac{\sigma_t \sqrt{2}}{\eta_t L} \right)^i \Gamma((i+1)/2). \tag{78}$$

Appendix D. Proof of equation (46)

The innermost integral of line (45) evaluates to

$$\int_{\eta_t L}^{\infty} \sup_{x_d : |x_d| \leq \eta_t L} f(w_d - x_d) \mathrm{d}w_d = \int_{\eta_t L}^{\infty} f(w_d - \eta_t L) \mathrm{d}w_d = \int_0^{\infty} f(w_d) \mathrm{d}w_d = \frac{1}{2}, \tag{79}$$

where the first equality follows from the monotonicity assumptions, the second from a change of variable, and the third from the symmetry assumption. Similarly, the innermost integral of line (44) evaluates to

$$\int_0^{\infty} \sup_{x_d : |x_d| \leq \eta_t L} f(w_d - x_d) \mathrm{d}w_d \tag{80}$$

$$= \int_0^{\eta_t L} \sup_{x_d : |x_d| \leq \eta_t L} f(w_d - x_d) \mathrm{d}w_d \mathrm{d}w^{d-1} + \int_{\eta_t L}^{\infty} \sup_{x_d : |x_d| \leq \eta_t L} f(w_d - x_d) \mathrm{d}w_d \tag{81}$$

$$= \eta_t L f(0) + \frac{1}{2}. \tag{82}$$

Combining equations (45), (79), and (82), we get

$$h_+(d) = \left(\eta_t L f(0) + \frac{1}{2}\right) \int\limits_{\overline{\mathcal{B}_\infty^{d-1}}(0,\eta_t L)\cap A_{d-1}} \prod_{i=1}^{d-1} \sup_{x_i:|x_i|\leq\eta_t L} f(w_i - x_i)\mathrm{d}w^{d-1} \tag{83}$$

$$+ \frac{1}{2} \int\limits_{\mathcal{B}_\infty^{d-1}(0,\eta_t L)\cap A_{d-1}} \prod_{i=1}^{d-1} \sup_{x_i:|x_i|\leq\eta_t L} f(w_i - x_i)\mathrm{d}w^{d-1} \tag{84}$$

$$= \left(\eta_t L f(0) + \frac{1}{2}\right) h_+(d-1) + \frac{1}{2}(\eta_t L f(0))^{d-1}, \tag{85}$$

where the second equality follows from the fact that f is maximized at 0, and $\mathcal{B}_\infty^{d-1}(0, \eta_t L) \cap A_{d-1}$ is a $(d-1)$-dimensional hypercube of side $\eta_t L$ (with volume $(\eta_t L)^{d-1}$). Now,

$$h(d) = 2^d h_+(d) = (1 + 2\eta_t L f(0))\, h(d-1) + (2\eta_t L f(0))^{d-1}. \tag{86}$$

Appendix E. Proof of Theorem 7

Consider any $f \in \mathcal{F}$, and let

$$f_+(x) = \begin{cases} f(x), & x \geq 0, \\ 0, & x < 0, \end{cases} \qquad \text{and} \qquad f_-(x) = \begin{cases} 0, & x \geq 0, \\ f(x), & x < 0. \end{cases} \tag{87}$$

Then

$$\mathbf{var}_f(X^2) = \int_{-\infty}^{+\infty} (f_-(x) + f_+(x))x^2 dx = \int_0^\infty 2f_+(x)x^2 dx, \tag{88}$$

where the second equality follows from the symmetry assumption. Note that $2f_+$ is a valid probability density over $[0, \infty)$, and let $X_+ \sim f_+$. Then, by previous equation,

$$\mathbf{var}_f(X^2) = \mathbf{E}_{(2f_+)}\left[X_+^2\right] = \int_0^\infty 2x\,(1 - \mathbf{Pr}(X_+ \leq x))\, dx \tag{89}$$

$$\geq \int_0^{1/(2f(0))} 2x\,(1 - 2xf(0))\, dx = \frac{1}{12f^2(0)}. \tag{90}$$

Hence,

$$f(0) \geq \frac{1}{2\sqrt{3}\sqrt{\mathbf{var}_f(X^2)}} \geq \frac{1}{2\sqrt{3}\sigma}, \tag{91}$$

which is achieved by the uniform distribution $\mathcal{U}(-\sigma\sqrt{3}, \sigma\sqrt{3})$. ■

Appendix F. Proof of Theorem 8

First, we show that the limit of the right-hand side of equation (47) is given by the right-hand side of equation (48). Note that

$$\frac{V_1(d, \eta_t L)}{(\sigma_t\sqrt{2})^d} = V_1\left(d, \frac{\eta_t L}{\sigma_t\sqrt{2}}\right) \xrightarrow{d\to\infty} 0. \tag{92}$$

On the other hand,

$$\lim_{d\to\infty}\sum_{i=0}^{d-1}\frac{(\sigma_t\eta_t L/\sqrt{2})^i}{i!}=\sum_{i=0}^{\infty}\frac{(\sigma_t\eta_t L/\sqrt{2})^i}{i!}=e^{\sigma_t\eta_t L/\sqrt{2}}. \tag{93}$$

Since T is finite, the limit and the sum are interchangeable, so that the above two equations yield the desired limit.

We now turn to the proof of inequality (47). For notational convenience, set $\lambda_t = \frac{\sigma_t}{\sqrt{2}}$ (so that $f_{t0}(x) = \frac{\lambda_t}{2}e^{-\lambda|x|}$ for all $x \in \mathbb{R}$) and $R_t = \eta_t L$. Since the noise satisfies the assumptions of Proposition 3, we get

$$\mathcal{L}\left(S{\to}W\right) \le \sum_{t=1}^{T}\log\left(f_t(0)V_1(d,R_t) + \int\limits_{\overline{\mathcal{B}_1}(0,R_t)}\sup_{x_t\in\mathcal{B}_1(0,R_t)} f_t(w_t-x_t)\mathrm{d}w_t\right) \tag{94}$$

$$= \sum_{t=1}^{T}\log\left(\frac{V_1(d,R_t)}{(\lambda_t/2)^d} + \int\limits_{\overline{\mathcal{B}_1}(0,R_t)}\sup_{x_t\in\mathcal{B}_1(0,R_t)}\left(\frac{\lambda_t}{2}\right)^d\exp\left\{-\lambda\|w_t-x_t\|_1\right\}\mathrm{d}w_t\right). \tag{95}$$

Recall $h(d,p,f_t,R_t)$ (cf. equation (17)) is defined to be the second term inside the log. Similarly to the strategy adopted in the proof of Theorem 6, we will derive a recurrence relation for h in terms of d, as such we will again suppress the dependence on p, f_t, and R_t in the notation, and write $h(d)$ only (and correspondingly $h_+(d)$).

Lemma 12 *Given $w \in \overline{\mathcal{B}_1^d}(0,R)\cap A_d$ (A_d defined in equation (18)),*

$$\inf_{x\in\mathcal{B}_1^d(0,R)}\|w-x\|_1 = \sum_{i=1}^{d} w_i - R. \tag{96}$$

Proof Since we are minimizing a continuous function over a compact set, then the infimum can be replaced with a minimum.

Claim: There exists a minimizer $x^\star$ such that for all i, $x_i^\star \le w_i$.

Proof of Claim: Consider any $x \in \mathcal{B}_1(0,R)$ such that there exists j satisfying $x_j > w_j$. Note that $w_j \ge 0$ by assumption. Now define $x' = (x_1,\ldots,x_{j-1},w_j,x_{j+1},\ldots,x_d)$. Then $\|x'\|_1 < \|x\|_1$ so that $x' \in \mathcal{B}_1(0,R)$. Moreover, $\|w-x'\|_1 \le \|w-x\|_1$ as desired. ■

Now,

$$\inf_{x\in\mathcal{B}_1^d(0,R)}\|w-x\|_1 = \inf_{\substack{x\in\mathcal{B}_1^d(0,R):\\ x_i\le w_i,\ \forall\ i}}\|w-x\|_1 = \inf_{\substack{x\in\mathcal{B}_1^d(0,R):\\ x_i\le w_i,\ \forall i}}\sum_{i=1}^{d}(w_i-x_i) = \sum_{i=1}^{d} w_i - R. \tag{97}$$

■

Given the above lemma, we will derive the recurrence relation by decomposing the integral over $\overline{\mathcal{B}_1^d}(0,R_t)$ into two disjoint subsets: 1) $w^{d-1} \notin \mathcal{B}_1^{d-1}(0,R_t)$, in which case w_d can take any value in $\mathbb{R}$, and 2) $w^{d-1} \in \mathcal{B}_1^{d-1}(0, R_t)$, in which case w_d must satisfy $|w_d| > R_t - \|w^{d-1}\|_1$.

$$h_+(d) = \int\limits_{\overline{\mathcal{B}_1^d}(0,R_t)\cap A_d} \sup_{x_t\in\mathcal{B}_1(0,R_t)} \left(\frac{\lambda_t}{2}\right)^d e^{-\lambda_t\left(\sum_{i=1}^d w_t - R_t\right)} \mathrm{d}w_t \tag{98}$$

$$= \int\limits_{\overline{\mathcal{B}_1^{d-1}}(0,R_t)\cap A_d} \left(\frac{\lambda_t}{2}\right)^{d-1} e^{-\lambda_t\left(\sum_{i=1}^{d-1} w_t - R_t\right)} \left(\int_0^\infty \frac{\lambda_t}{2} e^{-\lambda_t w_d} \mathrm{d}w_d\right) \mathrm{d}w^{d-1} \tag{99}$$

$$+ \int\limits_{\mathcal{B}_1^{d-1}(0,R_t)\cap A_d} \left(\frac{\lambda_t}{2}\right)^{d-1} e^{-\lambda_t\left(\sum_{i=1}^{d-1} w_t - R_t\right)} \left(\int_{R_t - \sum_{i=1}^{d-1} w_i}^\infty \frac{\lambda_t}{2} e^{-\lambda_t w_d} \mathrm{d}w_d\right) \mathrm{d}w^{d-1} \tag{100}$$

$$= \frac{1}{2} h_+(d-1) + \int\limits_{\mathcal{B}_1^{d-1}(0,R_t)\cap A_d} \left(\frac{\lambda_t}{2}\right)^{d-1} e^{-\lambda_t\left(\sum_{i=1}^{d-1} w_t - R_t\right)} \left(\frac{1}{2} e^{-\lambda_t(R_t - \sum_{i=1}^{d} w_i)}\right) \mathrm{d}w^{d-1} \tag{101}$$

$$= \frac{1}{2} h_+(d-1) + \frac{1}{2}\left(\frac{\lambda_t}{2}\right)^{d-1} \frac{V_1(d-1, R_t)}{2^{d-1}} \tag{102}$$

$$= \frac{1}{2} h_+(d-1) + \frac{1}{2}\left(\frac{\lambda_t R_t}{2}\right)^{d-1} \frac{1}{(d-1)!}. \tag{103}$$

Hence,

$$h(d) = 2^d h_+(d) = h(d-1) + \frac{(\lambda_t R_t)^{d-1}}{(d-1)!}. \tag{104}$$

It is easy check that $h(1) = 1$, and hence

$$h(d) = \sum_{i=0}^{d-1} \frac{(\lambda_t R_t)^i}{i!} \tag{105}$$

satisfies the base case and the recurrence relation. Re-substituting $\eta_t L$ and $\sigma_t/\sqrt{2}$ for R_t and λ_t, respectively, yields the desired result in equation (47).

Appendix G. Proof of Theorem 9

Let $R_t = \eta_t L$. Since the noise satisfies the assumptions of Proposition 3, we get

$$\mathcal{L}\left(S \to W\right) \leq \sum_{t=1}^T \log\left(f_t(0) V_1(d, R_t) + \int\limits_{\overline{\mathcal{B}_1}(0,R_t)} \sup_{x_t\in\mathcal{B}_1(0,R_t)} f_t(w_t - x_t) \mathrm{d}w_t\right) \tag{106}$$

$$= \sum_{t=1}^T \log\left(\frac{V_1(d, R_t)}{(2\pi\sigma^2)^{\frac{d}{2}}} + \int\limits_{\overline{\mathcal{B}_1}(0,R_t)} \sup_{x_t\in\mathcal{B}_1(0,R_t)} \frac{1}{(2\pi\sigma_t^2)^{\frac{d}{2}}} \exp\left\{-\frac{\|w_t - x_t\|_2^2}{2\sigma_t^2}\right\} \mathrm{d}w_t\right). \tag{107}$$

Consider

$$h_+(d) = \int\limits_{\overline{\mathcal{B}_1}(0,R_t)\cap A_d} \sup_{x_t \in \mathcal{B}_1(0,R_t)} \frac{1}{(2\pi\sigma_t^2)^{\frac{d}{2}}} \exp\left\{-\frac{\|w_t - x_t\|_2^2}{2\sigma_t^2}\right\} \mathrm{d}w_t. \tag{108}$$

First we solve $\inf\limits_{x_t \in \mathcal{B}_1(0,R_t)} \|w_t - x_t\|_2$. If $w_t \in A_d$, then the infimum is achieved for $x_t^\star \in A_d$ as well (one can simply flip the sign of any negative component, which cannot increase the distance). In the subspace A_d, the boundary of the L_1 ball is defined by the hyperplane $\sum_{i=1}^d x_{ti} = R_t$. As such, finding the minimum distance corresponds to projecting the point w to the given hyperplane:

$$\inf_{x_t \in \mathcal{B}_1(0,R_t)} \|w_t - x_t\|_2 = \min_{\substack{x_t \in \mathcal{B}_1(0,R_t)\cap A_d: \\ \sum_{i=1}^d x_i = R_t}} \|w_t - x_t\|_2 \geq \frac{\sum_{i=1}^d w_{ti} - R_t}{\sqrt{d}}. \tag{109}$$

Now,

$$h_+(d) \leq \int\limits_{\overline{\mathcal{B}_1}(0,R_t)\cap A_d} \frac{1}{(2\pi\sigma_t^2)^{\frac{d}{2}}} \exp\left\{-\frac{(\sum_{i=1}^d w_{ti} - R_t)^2}{2d\sigma_t^2}\right\} \mathrm{d}w_t. \tag{110}$$

For notational convenience, we drop the t subscript in the following. We perform a change of variable as follows: $\tilde{w}_d = \sum_{i=1}^d w_i$. Hence, for $w \notin \mathcal{B}_1(0,R)$, $\tilde{w}_d \geq R$. Since $w_d \geq 0$, then $\sum_{i=1}^{d-1} w_i \leq \tilde{w}_d$. For $x \in \mathbb{R}$, define $S(x) := \{w^{d-1} \in \mathbb{R}^{d-1} : \sum_{i=1}^{d-1} w_i \leq x\}$. Then,

$$\begin{aligned}
h_+(d) &= \int_R^\infty \int\limits_{S(\tilde{w}_d)} \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} e^{-\frac{(\tilde{w}_d - R)^2}{2d\sigma^2}} \mathrm{d}w^{d-1}\mathrm{d}w_d && (111)\\
&= \frac{1}{(2\pi\sigma_t^2)^{\frac{d}{2}}} \int_R^\infty e^{-\frac{(\tilde{w}_d - R)^2}{2d\sigma^2}} \left(\int\limits_{S(\tilde{w}_d)} \mathrm{d}w^{d-1} \right) \mathrm{d}w_d && (112)\\
&\overset{\text{(a)}}{=} \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}((d-1)!)} \int_R^\infty \tilde{w}_d^{d-1} e^{-\frac{(\tilde{w}_d - R)^2}{2d\sigma^2}} \mathrm{d}w_d && (113)\\
&\overset{\text{(b)}}{=} \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}((d-1)!)} R^{d-1}(\sigma\sqrt{2d}) \sum_{i=0}^{d-1} \left(\frac{\sigma\sqrt{2d}}{R}\right)^i \frac{\Gamma((i+1)/2)}{2}, && (114)
\end{aligned}$$

where (a) follows from the fact that the innermost integral corresponds to the volume of a scaled probability simplex (scaled by $\tilde{w}_d$), and (b) follows from the same computations as in Equations (73) to (76) (with $\tilde{\sigma} = \sigma\sqrt{d}$). Noting that $h(d) = 2^d h_+(d)$ yields the desired the term in Equation (50).