On Random Subset Generalization Error Bounds and the Stochastic Gradient Langevin Dynamics Algorithm

Borja Rodríguez-Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund

KTH – Royal Institute of Technology
ISE – Information Science and Engineering

March 20, 2021

Outline

What is the generalization error?

Standard learning scenario:

$$\mathcal{Z} \xrightarrow[N \text{ i.i.d}]{P_Z} S = (Z_1, \ldots, Z_N) \longrightarrow \boxed{P_{W|S}} \longrightarrow W$$

- The objective is to have a low **population risk**:

$$L_{P_Z}(W) \triangleq \mathbb{E}_{P_Z}[\ell(W, Z)].$$

- We don't know P_Z, so we observe the **empirical risk**:

$$L_S(W) \triangleq \frac{1}{N} \sum_{i=1}^{N} \ell(W, Z_i)$$

- Then we study the **generalization error**:

$$\mathrm{gen}(W, S) \triangleq L_{P_Z}(W) - L_S(W).$$

What is the generalization error?

An example, cat vs. dog image classification:

- $\mathcal{X}$ is the image space and f is the **true** mapping image $\rightarrow$ cat/dog.

$$\mathcal{X} \xrightarrow[N \text{ i.i.d}]{P_X} S = \big((X_1, f(X_1)), \ldots, (X_N, f(X_N))\big) \longrightarrow \boxed{\text{SGD}} \longrightarrow \begin{array}{l}\text{Neural Network}\\ \text{trained weights}\end{array}$$

- The objective is to have **low missclassification in the wild**.
- Since we don't know P_X nor f, we observe the **training missclassification**.
- Then, we study how it differs from **test missclassification** and **missclassification in production**.

Expected generalization error: standard setting

- For fixed w, it has mean 0 under P_S:

$$\mathbb{E}_{P_S}[\mathrm{gen}(w, S)] = L_{P_Z}(w) - \mathbb{E}_{P_S}[L_S(w)] = 0.$$

- The **expected generalization error** (EGE) is

$$\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W, S)] = \mathbb{E}_{P_W}[L_{P_Z}(W)] - \mathbb{E}_{P_{W,S}}[L_S(W)].$$

Expected generalization error: randomized-subsample setting

A more structured setting for sampling S **[Steinke and Zakynthinou]**:

$$\mathcal{Z} \xrightarrow[2N \text{ i.i.d}]{P_Z} \tilde{S} = \begin{bmatrix} Z_1 & Z_2 & \cdots & Z_N \\ Z_{N+1} & Z_{N+2} & \cdots & Z_{2N} \end{bmatrix} \longrightarrow S = (Z_1, Z_{N+2} \ldots, Z_{2N})$$

$$\{0,1\} \xrightarrow[N \text{ i.i.d}]{\text{Ber}(0.5)} U = (0, 1, \ldots, 1)$$

- Let $\bar{U} = (1 - U_1, \ldots, 1 - U_N)$.
- Then $S = \tilde{S}(U)$ and $\bar{S} = \tilde{S}(\bar{U}) = \tilde{S} \setminus S$.
- We also define the **empirical generalization error**:

$$\widehat{\text{gen}}(W, \tilde{S}, U) = L_{\bar{S}}(W) - L_S(W).$$

>Reasoning About Generalization via Conditional Mutual Information - T. Steinke and L. Zakynthinou - COLT (2020)

Expected generalization error: randomized-subsample setting

- For fixed $\tilde{s}$ and fixed w, it has mean 0 under $P_U = \mathrm{Ber}(0,5)^{\times N}$:

$$\mathbb{E}_{P_U}[\widehat{\mathrm{gen}}(w, \tilde{s}, U)] = \mathbb{E}_{P_U}[L_{\tilde{s}(\bar{U})}(W) - L_{\tilde{s}(U)}(w)] = 0.$$

- The **EGE** can be written as

$$\begin{aligned}\mathbb{E}_{P_{W,\tilde{S},U}}[\widehat{\mathrm{gen}}(W, \tilde{S}, U)] &= \mathbb{E}_{P_{W,\tilde{S},U}}[L_{\bar{S}}(W) - L_S(W)] \\ &= \mathbb{E}_{P_W}[L_{P_Z}(W)] - \mathbb{E}_{P_{W,S}}[L_S(W)] \\ &= \mathbb{E}_{P_{W,S}}[\mathrm{gen}(W, S)].\end{aligned}$$

Standard vs randomized-subsample setting: why do we care?

- **Xu and Raginsky** proved that, for σ^2-subgaussian losses:

$$\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| \leq \sqrt{\frac{2\sigma^2}{N} I(W;S)}.$$

  Problem: It can be infinite for e.g. deterministic algorithms.

- In the randomized-subsample setting we have that **[Steinke and Zakynthinou]**, for $[a,b]$-bounded losses:

$$\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| \leq \sqrt{\frac{2(b-a)^2}{N} I(W;U|\tilde{S})}.$$

  It is **always finite**: $I(W;U|\tilde{S}) \leq \log(2)N$.

> Information-theoretic analysis of generalization capability of learning algorithms - A. Xu and M. Raginsky - NeurIPS (2017)

>Reasoning About Generalization via Conditional Mutual Information - T. Steinke and L. Zakynthinou - COLT (2020)

Outline

Background: What is the generalization error?

- Standard Setting
- Randomized-subsample Setting

Part I: Deriving expected generalization error (EGE) bounds

- Deriving the main Lemma
- How to obtain bounds from this Lemma
- Limitations of this technique

Part II: EGE bounds for Stochastic Gradient Langevin Dynamics (SGLD)

- What is SGLD?
- From EGE general bounds to EGE SGLD boudns

Deriving the main Lemma

- **Main idea**: Abstract (Th.1 + Cor. 1 and Th.2 + Cor. 5) from **Hellström and Durisi** in a general lemma for arbitrary random variables.
- **Lemma**: If $f : \mathcal{X} \times \mathcal{Y}$ is either:
  1. Zero mean and σ^2-subgaussian under P_X for all $y \in \mathcal{Y}$, or
  2. Zero mean and σ^2-subgaussian under $P_X \times P_Y$.

  Then, for any Q_Y on $\mathcal{Y}$

$$\left|\mathbb{E}_{P_{X,Y}}[f(X,Y)]\right| \leq \sqrt{2\sigma^2 D_{\mathrm{KL}}(P_{X,Y}||Q_Y \times P_X)}.$$

> Generalization Bounds via Information Density and Conditional Information Density
- F. Hellström and Durisi - IEEE JSAIT (2020)

Deriving the main Lemma

- **Proof** (follows the rationale of **Hellström and Durisi**):
  - If $f(X, y)$ is σ^2-subgaussian under P_X for all $y \in \mathcal{Y}$ and $\mathbb{E}_{P_X}[f(X, y)] = 0$:

$$\mathbb{E}_{P_X}\left[\exp(\lambda f(X, y))\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

  - If we take the expectation w.r.t P_Y we obtain condition (2):

$$\mathbb{E}_{P_Y \times P_X}\left[\exp(\lambda f(X, Y))\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

  - We may take a **change of measure** and re-arrange the terms to get:

$$\mathbb{E}_{P_{X,Y}}\left[\exp\left(\lambda f(X, Y) - \frac{\lambda^2\sigma^2}{2} - \underbrace{\log\left(\frac{dP_{X,Y}}{d(P_X \times P_Y)}(X, Y)\right)}_{\iota(X,Y)\text{: information density}}\right)\right] \leq 1.$$

> Generalization Bounds via Information Density and Conditional Information Density
- F. Hellström and Durisi - IEEE JSAIT (2020)

Deriving the main Lemma

- **Proof** (cont.):
  - We obtained:

$$\mathbb{E}_{P_{X,Y}}\left[\exp\left(\lambda f(X,Y) - \frac{\lambda^2\sigma^2}{2} - \iota(X,Y)\right)\right] \leq 1.$$

  - Applying Jensen's inequality we have that

$$\exp\left(\lambda \mathbb{E}_{P_{X,Y}}[f(X,Y)] - \frac{\lambda^2\sigma^2}{2} - \underbrace{D_{\mathrm{KL}}(P_{X,Y}||P_Y \times P_X)}_{I(X;Y)\text{: mutual information}}\right) \leq 1.$$

  - Finally, solving for λ and using the **Golden formula** yields:

$$\left|\mathbb{E}_{P_{X,Y}}[f(X,Y)]\right| \leq \sqrt{2\sigma^2 D_{\mathrm{KL}}(P_{X,Y}||Q_Y \times P_X)}.$$

Obtaining bounds from the lemma

- We can recover Prop. 1 from **Bu et al.** If $\ell(w, Z)$ is σ^2-subgaussian for all $w \in \mathcal{W}$ under P_Z then

$$\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| \leq \frac{1}{N}\sum_{i=1}^{N}\sqrt{2\sigma^2 I(W;Z_i)} \leq \sqrt{\frac{2\sigma^2}{N} I(W;S)}.$$

- **Sketch of the proof**:
  - $f(X,Y) = \mathrm{gen}_i(W,Z_i) = \mathbb{E}_{P_Z}[\ell(W,Z)] - \ell(W,Z_i)$.
  - $P_X = P_{Z_i} = P_Z$ and $P_Y = Q_Y = P_W$.
  - Then, by the **Lemma**

$$\left|\mathbb{E}_{P_{W,Z_i}}[\mathrm{gen}_i(W,Z_i)]\right| \leq \sqrt{2\sigma^2 I(W;Z_i)}.$$

  - Finally, note that

$$\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| \leq \frac{1}{N}\sum_{i=1}^{N}\Big|\underbrace{\mathbb{E}_{P_W\times P_Z}[\ell(W,Z)] - \mathbb{E}_{P_{W,Z_i}}[\ell(W,Z_i)]}_{\mathbb{E}_{P_{W,Z_i}}[\mathrm{gen}_i(W,Z_i)]}\Big|$$

> Tightening Mutual Information Based Bounds on Generalization Error
-Y. Bu, S. Zou, and V. V. Veeravalli - IEEE JSAIT (2020)

Obtaining bounds from the lemma

- Using a similar strategy, we can extend Th. 2.4 from **[Negrea et al.]**
  - Let $\ell(w, Z)$ is σ^2-subgaussian for all $w \in \mathcal{W}$ under P_Z.
  - Let $J \subseteq [N]$ be an independent, uniformly random subset of indices of size M.
  - Let R be a random object only dependent of W.

  Then

$$\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| \leq \mathbb{E}_{P_{J,S_{J^c},R}}\left[\sqrt{\frac{2\sigma^2}{M} D_{\mathrm{KL}}(P_{W,S_J|S_{J^c},R} || Q_{W|S_{J^c},R} \times P_{S_J})}\right].$$

- **Main realization** for the proof:

$$\begin{aligned}\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| &= \left|\mathbb{E}_{P_{W,S,R,J}}[\mathrm{gen}(W,S)]\right| \\ &\leq \mathbb{E}_{P_{J,S_{J^c},R}}\left[\left|\mathbb{E}_{P_W}[L_{P_Z}(W)] - \mathbb{E}_{P_{W,S_J|S_{J^c},R}}[L_{S_J}(W)]\right|\right]\end{aligned}$$

- **Good** for SGLD.

> *Information-theoretic generalization bounds for SGLD via data-dependent estimates*
> *- J. Negrea, M. Haghifam, G.K. Dziugaite, A. Khisti and D. M. Roy - NeurIPS (2019)*

Obtaining bounds from the lemma

- **Analogously**, we get their counterparts in the randomized subsample setting, for $[a, b]$ bounded losses:
  - A **new** bound based on the conditional mutual information:

$$\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| \leq \frac{1}{N}\sum_{i=1}^{N}\sqrt{2(b-a)^2 I(W;U_i|\tilde{Z}_i,\tilde{Z}_{i+N})}.$$

  - Recover a combination of **Haghifam et al.'s** Theorem 2.1 and Lemma 3.6.

$$\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| \leq \mathbb{E}_{P_{J,\tilde{S},U_{J^c},R}}\left[\sqrt{\frac{2(b-a)^2}{M}D_{\mathrm{KL}}(P_{W,U_J|\tilde{S},U_{J^c},R}||Q_{W|\tilde{S},U_{J^c},R}\times P_{U_J})}\right].$$

> *Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms - M. Haghifam, J. Negrea, A. Khisti, D. M. Roy and G.K. Dziugaite - NeurIPS (2020)*

Limitation of the Lemma

- The lemma does not allow to pull out all expectations. E.g., for $M = 1$

$$\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| \leq \mathbb{E}_{P_{J,\tilde{S},\boldsymbol{U_{J^c}},R}}\left[\sqrt{2(b-a)^2 D_{\mathrm{KL}}\left(P_{W,U_J|\tilde{S},\boldsymbol{U_{J^c}},R}\middle|\middle|Q_{W|\tilde{S},U_{J^c},R}\times P_{U_J}\right)}\right].$$

vs.

$$\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| \leq \mathbb{E}_{P_{J,\tilde{S},\boldsymbol{U},R}}\left[\sqrt{2(b-a)^2 D_{\mathrm{KL}}\left(P_{W|\tilde{S},\boldsymbol{U},R}\middle|\middle|Q_{W|\tilde{S},U_{J^c},R}\right)}\right].$$

- We obtain the latter with a slight extension of **Haghifam et al.**

> *Sharpened generalization bounds based on conditional mutual informa-tion and an application to noisy, iterative algorithms - M. Haghifam, J. Negrea, A. Khisti, D. M. Roy and G.K. Dziugaite - NeurIPS (2020)*

Outline

Background: What is the generalization error?

- Standard Setting
- Randomized-subsample Setting

Part I: Deriving expected generalization error (EGE) bounds

- Deriving the main Lemma
- How to obtain bounds from this Lemma
- Limitations of this technique

Part II: EGE bounds for Stochastic Gradient Langevin Dynamics (SGLD)

- What is SGLD?
- From EGE general bounds to EGE SGLD boudns

SGLD

- Parametrized hypothesis W_θ. For us $W \equiv \theta \in \mathbb{R}^d$.
- Random inizialization $W_0 \sim \nu_W$.
- Iterative algorithm (T iterations):
  - η_t: learning rate at t
  - V_t: indices of the batch at t
  - $\epsilon_t \sim \mathcal{N}(0,1)$
  - Update rule:

$$W_t \leftarrow \underbrace{W_{t-1} - \eta_t \nabla_{W_{t-1}} L_{S_{V_t}}(W_{t-1})}_{SGD} + \sigma_t \epsilon_t$$

- Final hypothesis is W_T.

From EGE general bounds to EGE SGLD boudns

- Generic random-subset bound:

$$\left|\mathbb{E}_{P_{W,S}}[\mathrm{gen}(W,S)]\right| \leq \mathbb{E}_{P_{J,\tilde{S},U,R}}\left[\sqrt{2(b-a)^2 D_{\mathrm{KL}}\left(P_{W|\tilde{S},U,R}\middle|\middle|Q_{W|\tilde{S},U_{J^c},R}\right)}\right].$$

- For SGLD, we let R be the batch trajectory $V^T = (V_1, V_2, \ldots, V_T)$.

$$\left|\mathbb{E}_{P_{\boldsymbol{W_T},S}}[\mathrm{gen}(\boldsymbol{W_T},S)]\right| \leq \mathbb{E}_{P_{J,\tilde{S},U,\boldsymbol{V^T}}}\left[\sqrt{2(b-a)^2 \underbrace{D_{\mathrm{KL}}\left(P_{\boldsymbol{W_T}|\tilde{S},U,\boldsymbol{V^T}}\middle|\middle|Q_{\boldsymbol{W_T}|\tilde{S},U_{J^c},\boldsymbol{V^T}}\right)}_{\text{Let's upper bound this relative entropy}}}\right].$$

From EGE general bounds to EGE SGLD boudns

$$
\begin{aligned}
D_{\mathrm{KL}}\left(P_{W_T|\tilde{S},U,V^T}\middle|\middle|Q_{W_T|\tilde{S},U_{J^c},V^T}\right) &\overset{(a)}{\leq} D_{\mathrm{KL}}\left(P_{W^T|\tilde{S},U,V^T}\middle|\middle|Q_{W^T|\tilde{S},U_{J^c},V^T}\right) \\
&\overset{(b)}{\leq} \sum_{t=1}^{T} \mathbb{E}_{P_{W^{t-1}|\tilde{S},U,V^t}}\left[D_{\mathrm{KL}}\left(P_{W_t|W_{t-1},\tilde{S},U,V^t}\middle|\middle|Q_{W_t|W^{t-1},\tilde{S},U_{J^c},V^t}\right)\right] \\
&\overset{(c)}{\leq} \sum_{t\in\mathcal{T}_J(V^T)} \mathbb{E}_{P_{W^{t-1}|\tilde{S},U,V^t}}\Bigg[\underbrace{D_{\mathrm{KL}}\left(P_{W_t|W_{t-1},\tilde{S},U,V^t}\middle|\middle|Q_{W_t|W^{t-1},\tilde{S},U_{J^c},V^t}\right)}_{\text{Let's further upper bound this relative entropy}}\Bigg]
\end{aligned}
$$

- (a) Monotonicity of the relative entropy.
- (b) **Full chain rule** of relative entropy + **Markov property of SGLD**.
- (c) Restrict the sum to only non-zero terms (inspired by **Bu et al.**)
  - $\mathcal{T}_J(V^T)$: iterations where J was in the batches from V^T.

> Tightening Mutual Information Based Bounds on Generalization Error
-Y. Bu, S. Zou, and V. V. Veeravalli - IEEE JSAIT (2020)

From EGE general bounds to EGE SGLD boudns

- Note that $P_{W_t|W_{t-1},\tilde{S},U,V^t} = \mathcal{N}(\mu_{J,t}, \sigma_t^2)$, where

$$\mu_{J,t} = W_{t-1} - \frac{\eta_t}{|V_t|}\left(\sum_{\substack{i\in V_t\\ i\neq J}} \nabla\ell(W_{t-1}, Z_i) + (1-U_J)\underbrace{\nabla\ell(W_{t-1}, \tilde{Z}_J)}_{\text{selected if } U_J=0} + U_J\underbrace{\nabla\ell(W_{t-1}, \tilde{Z}_{J+N})}_{\text{selected if } U_J=1}\right)$$

- We can design $Q_{W_t|W^{t-1},\tilde{S},U_{J^c},V^t}$ as we want.
  - E.g., inspired by **Haghifam et al**, let $Q_{W_t|W^{t-1},\tilde{S},U_{J^c},V^t} = \mathcal{N}(\mu'_{J,t}, \sigma_t^2)$, where

$$\mu'_{J,t} = W_{t-1} - \frac{\eta_t}{|V_t|}\left(\sum_{\substack{i\in V_t\\ i\neq J}} \nabla\ell(W_{t-1}, Z_i) + (1-\pi_{J,t})\nabla\ell(W_{t-1}, \tilde{Z}_J) + \pi_{J,t}\nabla\ell(W_{t-1}, \tilde{Z}_{J+N})\right)$$

  - Here $\pi_{J,t}$ is an **estimation** of U_J.
  - **Log-likelihood ratio** estimator based on $W^t, \tilde{S}, U_{J^c}, V^{t-1}$.

> *Sharpened generalization bounds based on conditional mutual informa-tion and an application to noisy, iterative algorithms - M. Haghifam, J. Negrea, A. Khisti, D. M. Roy and G.K. Dziugaite - NeurIPS (2020)*

From EGE general bounds to EGE SGLD boudns

- Putting everything together + known relative entropy between Gaussians:

$$\left|\mathbb{E}_{P_{W,S}}\left[\mathrm{gen}(W,S)\right]\right| \leq \sqrt{2}(b-a)$$
$$\mathbb{E}_{P_{J,\tilde{S},U,V^T}}\left[\sqrt{\sum_{t\in\mathcal{T}_J(V^T)}\mathbb{E}_{P_{W^{t-1}|U,\tilde{S},V^{t-1}}}\left[\frac{\eta_t^2\|\zeta_{J,t}\|^2}{2\sigma_t^2|V_t|^2}(U_J-\pi_{J,t})^2\right]}\right],$$

- Where $\zeta_{J,t} = \nabla\ell(W_{t-1},\tilde{Z}_J) - \ell(W_{t-1},\tilde{Z}_{J+N})$ is the **two-sample incoherence**.
- Recovers Theorem 4.2 from **Haghifam et al** if $|V_t| = N$.

- And **more** in the **paper**
  - E.g. another choice of $Q_{W_t|W^{t-1},\tilde{S},U_{J^c},V^t}$ for tighter bounds...

> Sharpened generalization bounds based on conditional mutual informa-tion and an application to noisy, iterative algorithms - M. Haghifam, J. Negrea, A. Khisti, D. M. Roy and G.K. Dziugaite - NeurIPS (2020)