Information-theoretic Analysis of the Gibbs Algorithm: An Individual Sample Approach

Youheng Zhu[1] *Yuheng Bu*[2]

[1]Huazhong University of Science and Technology
[2]University of Florida

ITW 2024
Wednesday 27 November, 2024

Contents

Introduction

Learning-theoretic setups:

- W is the hypothesis, $w \in \mathcal{W} \Rightarrow$ the set of all candidate models.

Learning Algorithm as a Channel:

Introduction

Learning-theoretic setups:

- W is the hypothesis, $w \in \mathcal{W} \Rightarrow$ the set of all candidate models.
- $S = \{Z_i\}_{i=1}^n$ is the training data drawn from the distribution $\mu^{\otimes n}$

Learning Algorithm as a Channel:

Introduction

Learning-theoretic setups:

- W is the hypothesis, $w \in \mathcal{W} \Rightarrow$ the set of all candidate models.
- $S = \{Z_i\}_{i=1}^n$ is the training data drawn from the distribution $\mu^{\otimes n}$
- Loss function $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$

Learning Algorithm as a Channel:

Introduction

Learning-theoretic setups:

- W is the hypothesis, $w \in \mathcal{W} \Rightarrow$ the set of all candidate models.
- $S = \{Z_i\}_{i=1}^n$ is the training data drawn from the distribution $\mu^{\otimes n}$
- Loss function $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$

Learning Algorithm as a Channel:

- $P_{W|S}$ stochastically picks hypothesis given training data

$$S \longrightarrow \boxed{P_{W|S}} \longrightarrow W$$

Introduction

Learning-theoretic setups:

- W is the hypothesis, $w \in \mathcal{W} \Rightarrow$ the set of all candidate models.
- $S = \{Z_i\}_{i=1}^{n}$ is the training data drawn from the distribution $\mu^{\otimes n}$
- Loss function $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^{+}$

Learning Algorithm as a Channel:

- $P_{W|S}$ stochastically picks hypothesis given training data

$$\xrightarrow{S} \boxed{P_{W|S}} \xrightarrow{W}$$

- Randomness $\to$ SGD, initialization......

Introduction

Generalization Error:

- Empirical risk:

$$L_e(W, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i)$$

Introduction

Generalization Error:

- Empirical risk:

$$L_e(W,S)=\frac{1}{n}\sum_{i=1}^{n}\ell(W,Z_i)$$

- Population risk:

$$L_\mu(W)=\mathbb{E}_\mu[\ell(W,Z)]=\mathbb{E}_{\mu^{\otimes n}}[L_e(W,S)]$$

Introduction

Generalization Error:

- Empirical risk:

$$L_e(W, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i)$$

- Population risk:

$$L_\mu(W) = \mathbb{E}_\mu[\ell(W, Z)] = \mathbb{E}_{\mu^{\otimes n}}[L_e(W, S)]$$

- **"Generalization Error":** Difference between empirical risk and population risk

Introduction

Generalization Error:

- Empirical risk:

$$L_e(W, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i)$$

- Population risk:

$$L_\mu(W) = \mathbb{E}_\mu[\ell(W, Z)] = \mathbb{E}_{\mu^{\otimes n}}[L_e(W, S)]$$

- **"Generalization Error":** Difference between empirical risk and population risk
- $\text{gen}(P_{W|S}, P_S) = \mathbb{E}_{P_{W,S}}[L_\mu(W) - L_e(W, S)]$

Introduction

Generalization Error:

- Empirical risk:

$$L_e(W,S) = \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i)$$

- Population risk:

$$L_\mu(W) = \mathbb{E}_\mu[\ell(W,Z)] = \mathbb{E}_{\mu^{\otimes n}}[L_e(W,S)]$$

- **"Generalization Error":** Difference between empirical risk and population risk
- $\text{gen}(P_{W|S}, P_S) = \mathbb{E}_{P_{W,S}}[L_\mu(W) - L_e(W,S)]$
- Explanation: "Expected[Population risk - Empirical Risk]"

Introduction

Gibbs Algorithm

- Definition:

$$P^{[n]}_{W|S}(w|s) \triangleq \frac{\pi(w)e^{-\gamma L_e(w,s)}}{V_{L_e}(s,\gamma)}.$$

Why Gibbs Algorithm?

Introduction

Gibbs Algorithm

- Definition:

$$P_{W|S}^{[n]}(w|s) \triangleq \frac{\pi(w)e^{-\gamma L_e(w,s)}}{V_{L_e}(s,\gamma)}.$$

Why Gibbs Algorithm?

- Information-risk minimization (IRM)

$$P_{W|S}^{*} = \arg\min_{P_{W|S}} \left(\mathbb{E}[L_e(W,S)] + \frac{1}{\beta} I(W;S)\right)$$

Introduction

Gibbs Algorithm

- Definition:

$$P^{[n]}_{W|S}(w|s) \triangleq \frac{\pi(w)e^{-\gamma L_e(w,s)}}{V_{L_e}(s,\gamma)}.$$

Why Gibbs Algorithm?

- Information-risk minimization (IRM)

$$P^{*}_{W|S} = \arg\min_{P_{W|S}} \left(\mathbb{E}[L_e(W,S)] + \frac{1}{\beta} I(W;S) \right)$$

- Stochastic gradient Langevin dynamics (SGLD) 's limit behavior

Introduction

Gibbs Algorithm

- Definition:

$$P_{W|S}^{[n]}(w|s) \triangleq \frac{\pi(w)e^{-\gamma L_e(w,s)}}{V_{L_e}(s,\gamma)}.$$

Why Gibbs Algorithm?

- Information-risk minimization (IRM)

$$P_{W|S}^{*} = \arg\min_{P_{W|S}} \left(\mathbb{E}[L_e(W,S)] + \frac{1}{\beta} I(W;S) \right)$$

- Stochastic gradient Langevin dynamics (SGLD) 's limit behavior
- ...

Introduction

- Information-theoretic generalization error bound [Aolin Xu and Maxim Raginsky 2017]; [Gholamali Aminian et al. 2021]

$$|\mathrm{gen}(P_{W|S}, P_S)| \lesssim \sqrt{\frac{I(W;S)}{n}} \quad \text{Arbitrary algorithm}$$

$$\mathrm{gen}(P_{W|S}, P_S) = \frac{I_{SKL}(W;S)}{\gamma} \quad \text{Gibbs algorithm}$$

$$I_{SKL}(X;Y) \triangleq D_{KL}(P_{X,Y} \| P_X \otimes P_Y) + D_{KL}(P_X \otimes P_Y \| P_{X,Y}) = I(X;Y) + L(X;Y)$$

Introduction

- Information-theoretic generalization error bound [Aolin Xu and Maxim Raginsky 2017]; [Gholamali Aminian et al. 2021]

$$|\mathrm{gen}(P_{W|S}, P_S)| \lesssim \sqrt{\frac{I(W;S)}{n}} \quad \text{Arbitrary algorithm}$$

$$\mathrm{gen}(P_{W|S}, P_S) = \frac{I_{SKL}(W;S)}{\gamma} \quad \text{Gibbs algorithm}$$

$$I_{SKL}(X;Y) \triangleq D_{KL}(P_{X,Y}\|P_X \otimes P_Y) + D_{KL}(P_X \otimes P_Y\|P_{X,Y}) = I(X;Y) + L(X;Y)$$

- **Trade-Off:** Empirical risk vs Generalization

Introduction

- Information-theoretic generalization error bound [Aolin Xu and Maxim Raginsky 2017]; [Gholamali Aminian et al. 2021]

$$|\mathrm{gen}(P_{W|S}, P_S)| \lesssim \sqrt{\frac{I(W;S)}{n}} \quad \text{Arbitrary algorithm}$$

$$\mathrm{gen}(P_{W|S}, P_S) = \frac{I_{SKL}(W;S)}{\gamma} \quad \text{Gibbs algorithm}$$

$$I_{SKL}(X;Y) \triangleq D_{KL}(P_{X,Y} \| P_X \otimes P_Y) + D_{KL}(P_X \otimes P_Y \| P_{X,Y}) = I(X;Y) + L(X;Y)$$

- **Trade-Off:** Empirical risk vs Generalization
- **Simple Intuition:** Overfitting training data $\uparrow \Rightarrow$ Empirical risk $\downarrow \Rightarrow$ Information from training data $\uparrow \Rightarrow$ Generalization Error $\uparrow$

Introduction

$\mathbf{I(W;S)}$ vs $\mathbf{I(W;Z_i)}$

- Entire dataset to Individual sample [Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli 2020]

$$|\mathrm{gen}(P_{W|S}, P_S)| \lesssim \sqrt{\frac{I(W;S)}{n}} \quad \Rightarrow \quad |\mathrm{gen}(P_{W|S}, P_S)| \lesssim \frac{1}{n}\sum_{i=1}^{n}\sqrt{I(W;Z_i)}$$

$$\mathrm{gen}(P_{W|S}, P_S) = \frac{I_{SKL}(W;S)}{\gamma} \quad \Rightarrow \quad ?$$

Introduction

Our Contribution:

- Asymptotic equivalence between $I_{SKL}(W;S)$ and $I_{SKL}(W;Z_i)$

$$\mathrm{gen}(P_{W|S},P_S)=\frac{I_{SKL}(W;S)}{\gamma}\quad\Rightarrow\quad\mathrm{gen}(P_{W|S},P_S)\sim\frac{1}{\gamma}\cdot\sum_{i=1}^{n}I_{SKL}(W;Z_i)$$

- Rate and exact convergence speed of information:
  $I_{SKL}(W;S)\sim\sum_{i=1}^{n}I_{SKL}(W;Z_i)=\Theta(1/n)$
- Asymptotic equivalence between $I(W;S)$ and $L(W;S)$ $\Rightarrow$ a tighter generalization error bound.
- Illustrative Example: Mean Estimation

Non-asymptotic gap

Theorem

For joint distribution $P_{W,S}$ induced by the Gibbs algorithm, we have

$$\sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S) = \sum_{i=1}^{n}\left(\mathbb{E}_{P_{W,Z_i}}[J_i^{[n]}(W,Z_i)] - \mathbb{E}_{P_W\otimes P_{Z_i}}[J_i^{[n]}(W,Z_i)]\right),$$

where the Jensen gap $J_i^{[n]}(w,z_i)$ is defined as

$$\begin{aligned} J_i^{[n]}(w,z_i) \triangleq \log \int_{\mathcal{Z}^{n-1}} & P_{W|S}^{[n]}(w|z_i,z^{-i})d\mu^{n-1}(z^{-i}) \\ & - \int_{\mathcal{Z}^{n-1}} \log\left(P_{W|S}^{[n]}(w|z_i,z^{-i})\right)d\mu^{n-1}(z^{-i}), \end{aligned}$$

with $z^{-i} \triangleq \{z_1,\cdots,z_{i-1},z_{i+1},\cdots,z_n\}$.

Non-asymptotic gap

Remark

- Jensen gap $J_i^{[n]}(w, z_i)$ always non-negative.

Non-asymptotic gap

Remark

- Jensen gap $J_i^{[n]}(w, z_i)$ always non-negative.
- However, $\mathbb{E}_{P_{W,Z_i}}[J_i^{[n]}(W, Z_i)] - \mathbb{E}_{P_W \otimes P_{Z_i}}[J_i^{[n]}(W, Z_i)]$ can be either negative or positive.

Non-asymptotic gap

Remark

- Jensen gap $J_i^{[n]}(w, z_i)$ always non-negative.
- However, $\mathbb{E}_{P_{W,Z_i}}[J_i^{[n]}(W, Z_i)] - \mathbb{E}_{P_W \otimes P_{Z_i}}[J_i^{[n]}(W, Z_i)]$ can be either negative or positive.
- **The lack of Chaining rule** $\Rightarrow$ $I_{\mathrm{SKL}}(W; S)$ can be either larger or smaller than $\sum_{i=1}^{n} I_{\mathrm{SKL}}(W; Z_i)$.

  An example is provided in [Gholamali Aminian et al. 2021], Example 1.

Non-asymptotic gap

Remark

- Jensen gap $J_i^{[n]}(w, z_i)$ always non-negative.
- However, $\mathbb{E}_{P_{W,Z_i}}[J_i^{[n]}(W, Z_i)] - \mathbb{E}_{P_W \otimes P_{Z_i}}[J_i^{[n]}(W, Z_i)]$ can be either negative or positive.
- **The lack of Chaining rule** $\Rightarrow$ $I_{\mathrm{SKL}}(W; S)$ can be either larger or smaller than $\sum_{i=1}^{n} I_{\mathrm{SKL}}(W; Z_i)$.

  An example is provided in [Gholamali Aminian et al. 2021], Example 1.

- Characterizing the gap is difficult, turn to asymptotic as $n \to \infty$.

Asymptotic analysis

Limiting behavior of measure:

Lemma

- *$n \to \infty$, W and S tends to being independent:*

$$\lim_{n\to\infty}\left(\frac{dP^{[n]}_{W,Z^n}}{dP^{\infty}_{W}\otimes P_{Z^{\infty}}}\right)=1 \quad a.s.$$

Key: The Strong Law of Large Numbers: $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\ell(W,Z_i)=L_{\mu}(W) \quad a.s.$

Asymptotic analysis

Limiting behavior of measure:

Lemma

- $n \to \infty$, W *and* S *tends to being independent:*

$$\lim_{n\to\infty}\left(\frac{dP^{[n]}_{W,Z^n}}{dP^{\infty}_{W}\otimes P_{Z^\infty}}\right)=1 \quad a.s.$$

- $n \to \infty$, W *and* Z_i *tends to being independent:*

$$\liminf_{n\to\infty}\left(\frac{dP^{[n]}_{W,Z_i}}{dP^{\infty}_{W}\otimes P_{Z^\infty}}\right)=1 \quad a.s.$$

Key: The Strong Law of Large Numbers: $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\ell(W,Z_i)=L_\mu(W) \quad a.s.$

Asymptotic analysis

By exchanging limit and integral:

Theorem

If the loss function $\ell(w,z)$ is bounded, we have

$$I_{\mathrm{SKL}}(W;Z_i) \sim \frac{1}{n^2}\gamma^2\mathbb{E}_\mu\left[\mathbb{E}_W^\infty\left[(\ell(W,Z)-L_\mu(W))^2\right]-\mathbb{E}_W^\infty\left[(\ell(W,Z)-L_\mu(W))\right]^2\right].$$

In other words: $I_{SKL}(W;Z_i) = \Theta(1/n^2)$. The exact speed is given by the blue part.

Note that the $\mathbb{E}_\mu\left[\mathbb{E}_W^\infty\left[(\ell(W,Z)-L_\mu(W))^2\right]-\mathbb{E}_W^\infty\left[(\ell(W,Z)-L_\mu(W))\right]^2\right]$ is also associated with γ. So the speed isn't simply γ^2.

Asymptotic analysis

The gap: $\left| I_{\mathrm{SKL}}(W;S) - \sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) \right|$.

Theorem

For Gibbs algorithm with bounded loss function,

$$\left| I_{\mathrm{SKL}}(W;S) - \sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) \right| = o(1/n).$$

- Note that $I_{SKL}(W;Z_i) = \Theta(1/n^2)$, therefore the gap is order-wise negligible.

Asymptotic analysis

Sketch of Proof:

- Recall the non-asymptotic result

$$\sum_{i=1}^{n} I_{\mathrm{SKL}}(W; Z_i) - I_{\mathrm{SKL}}(W; S) = \sum_{i=1}^{n} \left(\mathbb{E}_{P_{W,Z_i}}[J_i^{[n]}(W, Z_i)] - \mathbb{E}_{P_W \otimes P_{Z_i}}[J_i^{[n]}(W, Z_i)] \right)$$

Asymptotic analysis

Sketch of Proof:

- Recall the non-asymptotic result

$$\sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S) = \sum_{i=1}^{n}\left(\mathbb{E}_{P_{W,Z_i}}[J_i^{[n]}(W,Z_i)] - \mathbb{E}_{P_W\otimes P_{Z_i}}[J_i^{[n]}(W,Z_i)]\right)$$

- **lemma:** z_i in $J_i^{[n]}(w,z_i)$ is order-wise neglectable, so $\lim_{n\to\infty} n\cdot J_i^{[n]}(w,z_i) = \lim_{n\to\infty} n\cdot \hat{J}^{[n]}(w) = K_0(w)$.

$$\begin{aligned}
&\lim_{n\to\infty} n\cdot\left(\sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S)\right) \\
=& \lim_{n\to\infty}\int_{W,Z^\infty} n\cdot J_i^{[n]}(w,z_i)\cdot n\cdot\left(1-\frac{dP_W^{[n]}\otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}\right)\left(\frac{dP_{W,Z_i}^{[n]}}{dP_W^\infty\otimes P_{Z^\infty}}\right)dP_W^\infty\otimes P_{Z^\infty} \\
=& \int_{W,Z_i} K_0(w)\Big(-(\gamma(\ell(w,z_i)-L_\mu(w)))+\mathbb{E}_W^\infty[\gamma(\ell(W,z_i)-L_\mu(W))]\Big)\cdot 1\cdot dP_W^\infty\otimes P_{Z_i} \\
=& 0.
\end{aligned}$$

Asymptotic analysis

Sketch of Proof:

- Recall the non-asymptotic result

$$\sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S) = \sum_{i=1}^{n}\left(\mathbb{E}_{P_{W,Z_i}}[J_i^{[n]}(W,Z_i)] - \mathbb{E}_{P_W\otimes P_{Z_i}}[J_i^{[n]}(W,Z_i)]\right)$$

- **lemma:** z_i in $J_i^{[n]}(w,z_i)$ is order-wise neglectable, so $\lim_{n\to\infty} n\cdot J_i^{[n]}(w,z_i) = \lim_{n\to\infty} n\cdot \hat{J}^{[n]}(w) = K_0(w)$.

$$\begin{aligned}
&\lim_{n\to\infty} n\cdot\left(\sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S)\right)\\
=&\lim_{n\to\infty}\int_{W,Z^\infty} n\cdot J_i^{[n]}(w,z_i)\cdot n\cdot\left(1-\frac{dP_W^{[n]}\otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}\right)\left(\frac{dP_{W,Z_i}^{[n]}}{dP_W^\infty\otimes P_{Z^\infty}}\right)dP_W^\infty\otimes P_{Z^\infty}\\
=&\int_{W,Z_i} K_0(w)\left(-(\gamma(\ell(w,z_i)-L_\mu(w))) + \mathbb{E}_W^\infty[\gamma(\ell(W,z_i)-L_\mu(W))]\right)\cdot 1\cdot dP_W^\infty\otimes P_{Z_i}\\
=&\,0.
\end{aligned}$$

Asymptotic analysis

How do $I(W;S)$ and $L(W;S)$ scale respectively?

- **Existing result for Gaussian channel:**
  $L(W;S) > I(W;S)$ for Gaussian channel $P_{W|S}$.

Asymptotic analysis

How do $I(W;S)$ and $L(W;S)$ scale respectively?

- **Existing result for Gaussian channel:**
  $L(W;S) > I(W;S)$ for Gaussian channel $P_{W|S}$.

Asymptotic analysis

How do $I(W;S)$ and $L(W;S)$ scale respectively?

- **Existing result for Gaussian channel:**
  $L(W;S) > I(W;S)$ for Gaussian channel $P_{W|S}$.

Theorem

Mutual Information and Lautum Information are asymptotically equivalent.

$$I(W;S) \sim L(W;S) \sim \frac{1}{2} I_{SKL}(W;S)$$

Comparison to Algorithm Stability

- Results from algorithm stability [Maxim Raginsky et al. 2016]

$$\ell(w, z) \in [0, 1], \quad |\mathrm{gen}(P_{W|S}^{[n]}, P_S)| \leq \frac{\gamma}{2n}$$

Comparison to Algorithm Stability

- Results from algorithm stability [Maxim Raginsky et al. 2016]

$$\ell(w,z) \in [0,1], \quad |\text{gen}(P^{[n]}_{W|S}, P_S)| \leq \frac{\gamma}{2n}$$

- **Coin-tossing example:** $w \in \{0,1\}$ and $z \in \{0,1\}$, $\ell(w,z) = \mathbb{1}_{w=z}$ is a zero-one loss, and $\pi(w)$ is uniform over $\{0,1\}$.

  Our previous result shows that:

$$\text{gen}(P^{[n]}_{W|S}, P_S) \sim \frac{\gamma}{4n}$$

Comparison to Algorithm Stability

- Results from algorithm stability [Maxim Raginsky et al. 2016]

$$\ell(w,z) \in [0,1], \quad |\mathrm{gen}(P_{W|S}^{[n]}, P_S)| \le \frac{\gamma}{2n}$$

- **Coin-tossing example:** $w \in \{0,1\}$ and $z \in \{0,1\}$, $\ell(w,z) = \mathbb{1}_{w=z}$ is a zero-one loss, and $\pi(w)$ is uniform over $\{0,1\}$.

  Our previous result shows that:

$$\mathrm{gen}(P_{W|S}^{[n]}, P_S) \sim \frac{\gamma}{4n}$$

- As a development of the previous theorem $(I(W;S) \sim L(W;S))$, we show that

Comparison to Algorithm Stability

- Results from algorithm stability [Maxim Raginsky et al. 2016]

$$\ell(w,z) \in [0,1], \quad |\mathrm{gen}(P^{[n]}_{W|S}, P_S)| \le \frac{\gamma}{2n}$$

- **Coin-tossing example:** $w \in \{0,1\}$ and $z \in \{0,1\}$, $\ell(w,z) = \mathbb{1}_{w=z}$ is a zero-one loss, and $\pi(w)$ is uniform over $\{0,1\}$.

  Our previous result shows that:

$$\mathrm{gen}(P^{[n]}_{W|S}, P_S) \sim \frac{\gamma}{4n}$$

- As a development of the previous theorem ($I(W;S) \sim L(W;S)$), we show that

Theorem

For $\ell(w,z) \in [0,1]$, $\forall \delta > 0$, there exist an $N \in \mathbb{N}^+$ such that $\forall n > N$,

$$0 \le \mathrm{gen}(P^{[n]}_{W|S}, P_S) \le \frac{\gamma}{(4-\delta)n}.$$

Mean Estimation

We consider the problem of learning the means of the distribution μ. For simplicity, consider 1-dimensional case.

- $S = \{Z_i\}_{i=1}^n$, $Z_i \sim \mu = \mathcal{N}(0, (\frac{1}{\sqrt{2\beta}})^2)$, i.i.d.

Then, for a Gibbs algorithm with inverse temperature γ:

Mean Estimation

We consider the problem of learning the means of the distribution μ. For simplicity, consider 1-dimensional case.

- $S = \{Z_i\}_{i=1}^n$, $Z_i \sim \mu = \mathcal{N}(0, (\frac{1}{\sqrt{2\beta}})^2)$, i.i.d.
- Square error $\ell(w, Z) \triangleq \|w - Z\|_2^2 = (w - Z)^2$. **Unbounded!**

Then, for a Gibbs algorithm with inverse temperature γ:

Mean Estimation

We consider the problem of learning the means of the distribution μ. For simplicity, consider 1-dimensional case.

- $S = \{Z_i\}_{i=1}^n$, $Z_i \sim \mu = \mathcal{N}(0, (\frac{1}{\sqrt{2\beta}})^2)$, i.i.d.
- Square error $\ell(w, Z) \triangleq \|w - Z\|_2^2 = (w - Z)^2$. **Unbounded!**
- Prior distribution $\pi(w) = \frac{1}{\sqrt{\pi}} \exp(-w^2)$.

Then, for a Gibbs algorithm with inverse temperature γ:

Mean Estimation

We consider the problem of learning the means of the distribution μ. For simplicity, consider 1-dimensional case.

- $S = \{Z_i\}_{i=1}^n$, $Z_i \sim \mu = \mathcal{N}(0, (\frac{1}{\sqrt{2\beta}})^2)$, i.i.d.
- Square error $\ell(w, Z) \triangleq \|w - Z\|_2^2 = (w - Z)^2$. **Unbounded!**
- Prior distribution $\pi(w) = \frac{1}{\sqrt{\pi}} \exp(-w^2)$.

Then, for a Gibbs algorithm with inverse temperature γ:

- $\gamma \mathrm{gen}(P_{W|S}, \mu) = I_{\mathrm{SKL}}(W; S) = \frac{\gamma^2}{n\beta(1+\gamma)}$

Mean Estimation

We consider the problem of learning the means of the distribution μ. For simplicity, consider 1-dimensional case.

- $S = \{Z_i\}_{i=1}^n$, $Z_i \sim \mu = \mathcal{N}(0, (\frac{1}{\sqrt{2\beta}})^2)$, i.i.d.
- Square error $\ell(w, Z) \triangleq \|w - Z\|_2^2 = (w - Z)^2$. **Unbounded!**
- Prior distribution $\pi(w) = \frac{1}{\sqrt{\pi}} \exp(-w^2)$.

Then, for a Gibbs algorithm with inverse temperature γ:

- $\gamma \mathrm{gen}(P_{W|S}, \mu) = I_{\mathrm{SKL}}(W; S) = \frac{\gamma^2}{n\beta(1+\gamma)}$
- $I_{\mathrm{SKL}}(W; Z_i) = \frac{\gamma^2}{n^2\beta(1+\gamma)+\gamma^2(n-1)}$

Mean Estimation

We consider the problem of learning the means of the distribution μ. For simplicity, consider 1-dimensional case.

- $S = \{Z_i\}_{i=1}^n$, $Z_i \sim \mu = \mathcal{N}(0, (\frac{1}{\sqrt{2\beta}})^2)$, i.i.d.
- Square error $\ell(w, Z) \triangleq \|w - Z\|_2^2 = (w - Z)^2$. **Unbounded!**
- Prior distribution $\pi(w) = \frac{1}{\sqrt{\pi}} \exp(-w^2)$.

Then, for a Gibbs algorithm with inverse temperature γ:

- $\gamma \mathrm{gen}(P_{W|S}, \mu) = I_{\mathrm{SKL}}(W; S) = \frac{\gamma^2}{n\beta(1+\gamma)}$
- $I_{\mathrm{SKL}}(W; Z_i) = \frac{\gamma^2}{n^2\beta(1+\gamma)+\gamma^2(n-1)}$
- $\sum_{i=1}^n I_{\mathrm{SKL}}(W; Z_i) - I_{\mathrm{SKL}}(W; S) = \Theta(1/n^2) = o(1/n)$

Mean Estimation

We consider the problem of learning the means of the distribution μ. For simplicity, consider 1-dimensional case.

- $S = \{Z_i\}_{i=1}^n$, $Z_i \sim \mu = \mathcal{N}(0, (\frac{1}{\sqrt{2\beta}})^2)$, i.i.d.
- Square error $\ell(w, Z) \triangleq \|w - Z\|_2^2 = (w - Z)^2$. **Unbounded!**
- Prior distribution $\pi(w) = \frac{1}{\sqrt{\pi}} \exp(-w^2)$.

Then, for a Gibbs algorithm with inverse temperature γ:

- $\gamma \mathrm{gen}(P_{W|S}, \mu) = I_{\mathrm{SKL}}(W; S) = \frac{\gamma^2}{n\beta(1+\gamma)}$
- $I_{\mathrm{SKL}}(W; Z_i) = \frac{\gamma^2}{n^2\beta(1+\gamma)+\gamma^2(n-1)}$
- $\sum_{i=1}^n I_{\mathrm{SKL}}(W; Z_i) - I_{\mathrm{SKL}}(W; S) = \Theta(1/n^2) = o(1/n)$
- $I(W; S) = \frac{1}{2} \log\left(1 + \frac{\gamma^2}{n^2(1+\gamma)\beta+(n-1)\gamma^2}\right) \sim \frac{1}{2} I_{SKL}(W; S)$

Summary

Conclusion

- Asymptotic equivalence between $I_{SKL}(W;S)$ and $I_{SKL}(W;Z_i)$
- Rate and exact convergence speed of information:
  $I_{SKL}(W;S) \sim \sum_{i=1}^{n} I_{SKL}(W;Z_i) = \Theta(1/n)$
- Asymptotic equivalence between $I(W;S)$ and $L(W;S)$ $\Rightarrow$ a tighter generalization error bound.
- Example: Mean Estimation

Future Works

- **Conjecture:** The results hold for not only bounded loss function, but for loss function with a light tail distribution (e.g. σ-subgaussian)
- The relation between sample size n and inverse temperature γ.

Reference

[Ami+21] Gholamali Aminian et al. "An exact characterization of the generalization error for the Gibbs algorithm". In: *Proc. Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021), pp. 8106–8118.

[BZV20] Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. "Tightening mutual information-based bounds on generalization error". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 121–130.

[Rag+16] Maxim Raginsky et al. "Information-theoretic analysis of stability and bias of learning algorithms". In: *2016 IEEE Information Theory Workshop (ITW)*. IEEE. 2016, pp. 26–30.

[XR17] Aolin Xu and Maxim Raginsky. "Information-theoretic analysis of generalization capability of learning algorithms". In: *Advances in neural information processing systems* 30 (2017).

[Zou+24] Xinying Zou et al. "The Worst-Case Data-Generating Probability Measure in Statistical Learning". In: *IEEE Journal on Selected Areas in Information Theory* 5 (2024), pp. 175–189.

Thanks for listening.

Non-asymptotic gap

WCDG Distribution [Xinying Zou et al. 2024]:

- WCDG distribution $P_{\hat{S}|\Theta=\theta}^{(P_0,\beta)}$ is defined as:

$$\frac{dP_{\hat{S}|\Theta=\theta}^{(P_0,\beta)}}{dP_0} = \exp\left(\frac{1}{\beta}\ell(\theta,s) - \log\int \exp\left(\frac{1}{\beta}\ell(\theta,s)\right) dP_0(s)\right)$$

Non-asymptotic gap

WCDG Distribution [Xinying Zou et al. 2024]:

- WCDG distribution $P^{(P_0,\beta)}_{\hat{S}|\Theta=\theta}$ is defined as:

$$\frac{dP^{(P_0,\beta)}_{\hat{S}|\Theta=\theta}}{dP_0} = \exp\left(\frac{1}{\beta}\ell(\theta, s) - \log \int \exp\left(\frac{1}{\beta}\ell(\theta, s)\right) dP_0(s)\right)$$

- Represents worst-case distribution maximizing expected loss, given a reference P_0

Non-asymptotic gap

WCDG Distribution [Xinying Zou et al. 2024]:

- WCDG distribution $P^{(P_0,\beta)}_{\hat{S}|\Theta=\theta}$ is defined as:

$$\frac{dP^{(P_0,\beta)}_{\hat{S}|\Theta=\theta}}{dP_0} = \exp\left(\frac{1}{\beta}\ell(\theta, s) - \log \int \exp\left(\frac{1}{\beta}\ell(\theta, s)\right) dP_0(s)\right)$$

- Represents worst-case distribution maximizing expected loss, given a reference P_0
- Connected to optimization problem:

$$\max_{P \ll P_0} \int \ell(\theta, s) dP(s) \quad \text{s.t.} \quad D(P \| P_0) \leq \alpha$$

Non-asymptotic gap

Worst-case Data Generation (WCDG):

- Assign $\theta = w, z_i,\ s = z^{-i}$, and define:

$$\ell(\theta, s) = L_e(w, z_i, z^{-i}) - \frac{1}{\gamma} \log V_{L_e}(z_i, z^{-i}, \gamma)$$

Non-asymptotic gap

Worst-case Data Generation (WCDG):

- Assign $\theta = w, z_i,\ s = z^{-i}$, and define:

$$\ell(\theta, s) = L_e(w, z_i, z^{-i}) - \frac{1}{\gamma}\log V_{L_e}(z_i, z^{-i}, \gamma)$$

- Leads to upper bound for generalization error:

$$\sum_{i=1}^{n}\left(I_{\mathrm{SKL}}(W; Z_i) + D(P_W \otimes P_S \| P^{(\mu^{n-1}, \frac{1}{\gamma})}_{\hat{Z}^{n-1}, Z_i, W})\right)$$
$$\geq I_{\mathrm{SKL}}(W; S) = \gamma \mathrm{gen}(P^{\gamma}_{W|S}, P_S)$$

Asymptotic analysis

Sketch of Proof: By the Strong Law of Large Numbers

- Probability space $(\Omega, \mathcal{A}, P_\Omega)$ where $\{Z_i\}_{i=1}^n$ are i.i.d. random variables (not necessarily taking value in $\mathbb{R}$).

Asymptotic analysis

Sketch of Proof: By the Strong Law of Large Numbers

- Probability space $(\Omega, \mathcal{A}, P_\Omega)$ where $\{Z_i\}_{i=1}^n$ are i.i.d. random variables (not necessarily taking value in $\mathbb{R}$).
- $(\Omega, \mathcal{A}, P_\Omega)$ push forward to $(\mathcal{Z}^\infty, \mathcal{F}^\infty, \{\mathcal{F}^n\}, P_{Z^\infty})$.

Asymptotic analysis

Sketch of Proof: By the Strong Law of Large Numbers

- Probability space $(\Omega, \mathcal{A}, P_\Omega)$ where $\{Z_i\}_{i=1}^n$ are i.i.d. random variables (not necessarily taking value in $\mathbb{R}$).
- $(\Omega, \mathcal{A}, P_\Omega)$ push forward to $(\mathcal{Z}^\infty, \mathcal{F}^\infty, \{\mathcal{F}^n\}, P_{Z^\infty})$.
- Overall space of interest: Push forward space product with $(\mathcal{W}, \mathcal{B}, P_W^\infty)$ gets $(\mathcal{W} \times \mathcal{Z}^\infty, \mathcal{B} \times \mathcal{F}^\infty, P_W^\infty \otimes P_{Z^\infty})$.

Asymptotic analysis

Sketch of Proof: By the Strong Law of Large Numbers

- Probability space $(\Omega, \mathcal{A}, P_\Omega)$ where $\{Z_i\}_{i=1}^n$ are i.i.d. random variables (not necessarily taking value in $\mathbb{R}$).
- $(\Omega, \mathcal{A}, P_\Omega)$ push forward to $(\mathcal{Z}^\infty, \mathcal{F}^\infty, \{\mathcal{F}^n\}, P_{Z^\infty})$.
- Overall space of interest: Push forward space product with $(\mathcal{W}, \mathcal{B}, P_W^\infty)$ gets $(\mathcal{W} \times \mathcal{Z}^\infty, \mathcal{B} \times \mathcal{F}^\infty, P_W^\infty \otimes P_{Z^\infty})$.
- Every distribution $P_{W,Z^n}^{[n]}$ defined by Gibbs algorithm (as a markov kernel) on $(\mathcal{W} \times \mathcal{Z}^\infty, \mathcal{B} \times \mathcal{F}^n)$, $P_{W,Z_i}^{[n]}$ as its marginalization.

Asymptotic analysis

Sketch of Proof: By the Strong Law of Large Numbers

- Probability space $(\Omega, \mathcal{A}, P_\Omega)$ where $\{Z_i\}_{i=1}^n$ are i.i.d. random variables (not necessarily taking value in $\mathbb{R}$).
- $(\Omega, \mathcal{A}, P_\Omega)$ push forward to $(\mathcal{Z}^\infty, \mathcal{F}^\infty, \{\mathcal{F}^n\}, P_{Z^\infty})$.
- Overall space of interest: Push forward space product with $(\mathcal{W}, \mathcal{B}, P_W^\infty)$ gets $(\mathcal{W} \times \mathcal{Z}^\infty, \mathcal{B} \times \mathcal{F}^\infty, P_W^\infty \otimes P_{Z^\infty})$.
- Every distribution $P_{W,Z^n}^{[n]}$ defined by Gibbs algorithm (as a markov kernel) on $(\mathcal{W} \times \mathcal{Z}^\infty, \mathcal{B} \times \mathcal{F}^n)$, $P_{W,Z_i}^{[n]}$ as its marginalization.
- The Strong Law of Large Numbers indicates:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i) = L_\mu(W) \quad a.s.$$

w.r.t. $P_W^\infty \otimes P_{Z^\infty}$

Asymptotic proof

- By Fatou's lemma

$$
\begin{aligned}
\lim_{n\to\infty}\left(\frac{dP^{[n]}_{W,Z_i}}{dP^{\infty}_{W}\otimes P_{Z^\infty}}\right) &= \lim_{n\to\infty}\left(\int_{\mathcal{Z}^{n-1}} P^{[n]}_{W|Z_i,z^{-i}}d\mu^{n-1}(z^{-i})\right)\cdot\frac{d\mathrm{Leb}}{dP^{\infty}_{W}} \\
&\geq \left(\int_{\mathrm{Z}^\infty}\lim_{n\to\infty} P^{[n]}_{W|Z_i,z^{-i}}dP_{Z^\infty}(z^\infty)\right)\cdot\frac{d\mathrm{Leb}}{dP^{\infty}_{W}} \\
&= 1 \quad \text{Strong Law of Large Numbers}
\end{aligned}
$$

Asymptotic proof

- By Fatou's lemma

$$\underline{\lim}_{n\to\infty}\left(\frac{dP^{[n]}_{W,Z_i}}{dP^{\infty}_{W}\otimes P_{Z^\infty}}\right) = \underline{\lim}_{n\to\infty}\left(\int_{\mathcal{Z}^{n-1}} P^{[n]}_{W|Z_i,z^{-i}} d\mu^{n-1}(z^{-i})\right)\cdot\frac{d\mathrm{Leb}}{dP^{\infty}_{W}}$$
$$\geq \left(\int_{\mathrm{Z}^\infty}\underline{\lim}_{n\to\infty} P^{[n]}_{W|Z_i,z^{-i}} dP_{Z^\infty}(z^\infty)\right)\cdot\frac{d\mathrm{Leb}}{dP^{\infty}_{W}}$$
$$= 1 \quad \text{Strong Law of Large Numbers}$$

- Suppose $\Pr(\liminf_{n\to\infty}(dP^{[n]}_{W,Z_i}/dP^{\infty}_{W}\otimes P_{Z^\infty}) > 1) > 0$, again applying Fatou's lemma from another direction

$$1 = \underline{\lim}_{n\to\infty}\int_{\mathcal{W}\times\mathcal{Z}^\infty}\left(\frac{dP^{[n]}_{W,Z_i}}{dP^{\infty}_{W}\otimes P_{Z^\infty}}\right)dP^{\infty}_{W}\otimes P_{Z^\infty}$$
$$\geq \int_{\mathcal{W}\times\mathcal{Z}^\infty}\underline{\lim}_{n\to\infty}\left(\frac{dP^{[n]}_{W,Z_i}}{dP^{\infty}_{W}\otimes P_{Z^\infty}}\right)dP^{\infty}_{W}\otimes P_{Z^\infty} > 1 \quad ✗$$

So $\Pr(\liminf_{n\to\infty}(dP^{[n]}_{W,Z_i}/dP^{\infty}_{W}\otimes P_{Z^\infty}) = 1) = 1$

Bound by Stability

Condition: $\exp(-2\beta/n) < \mathrm{d}A_s^\beta/\mathrm{d}A_{s'}^\beta < \exp(2\beta/n)$

Target: Upper bound $D(A_s^\beta \| A_{s'}^\beta)$ using Hoeffding's lemma.

Proof: From the definition,

$$D(A_s^\beta \| A_{s'}^\beta) = \mathbb{E}_{A_s^\beta}\left[\log \frac{\mathrm{d}A_s^\beta}{\mathrm{d}A_{s'}^\beta}\right] \leq \frac{2\beta}{n}$$

which is the suboptimal bound. Using Hoeffding's lemma, for any r.v. $X \in [a, b]$,

$$\mathbb{E}[e^X] \leq \exp\left(\mathbb{E}[X] + \frac{(b-a)^2}{8}\right)$$

Letting $X = -\log \mathrm{d}A_s^\beta/\mathrm{d}A_{s'}^\beta$, we get

$$1 \leq \exp\left(-D(A_s^\beta \| A_{s'}^\beta) + \frac{(\frac{4\beta}{n})^2}{8}\right)$$

$$\Rightarrow \quad D(A_s^\beta \| A_{s'}^\beta) \leq \frac{2\beta^2}{n^2}$$