Tutorial
Characterizing the Generalization Error of Machine Learning Algorithms via Information Measures

Gholamali Aminian, Yuheng Bu, Iñaki Esnaola, and Samir M. Perlaza

2024 IEEE Information Theory Workshop

The 24th of November, 2024
Shenzhen, China

Slides for Part II

Table of Contents (Part II)

- KL divergence: $\mathrm{KL}(P\|Q) \triangleq \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP$
- Symmetrized KL divergence (Jeffrey's divergence)

$$D_{\mathrm{SKL}}(P\|Q) \triangleq \mathrm{KL}(P\|Q) + \mathrm{KL}(Q\|P).$$

- Mutual information: $\mathrm{I}(X;Y) \triangleq \mathrm{KL}(P_{X,Y}\|P_X \otimes P_Y)$
- Lautum information [Palomar and Verdú, 2008]: $L(X;Y) \triangleq \mathrm{KL}(P_X \otimes P_Y\|P_{X,Y})$
- Symmetrized KL information [Aminian et al., 2015]:

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

Information-theoretic Generalization Bounds

Lemma ([Xu and Raginsky, 2017])

Suppose $\ell(w, Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all $w \in \mathcal{W}$, then

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{2\sigma^2}{n}\mathrm{I}(S; W)}.$$

- Depends on every ingredient in the supervised learning problem
- Reducing dependence between W and S leads to better generalization bound
- This bound is only tight if $\mathrm{I}(S; W) = 0$ and $\mathrm{gen}(\mu, P_{W|S}) = 0$
- Multiple techniques to improve this result, including ISMI [Bu et al., 2020], CMI [Steinke and Zakynthinou, 2020], f-CMI [Harutyunyan et al., 2021], ΔL-CMI [Wang and Mao, 2023]

Table of Contents (Part II)

Regularized ERM problem

- How can we use this result to develop a better learning algorithm?
- Regularizing mutual information $\mathrm{I}(S;W)$ during ERM

$$P^{\star}_{W|S} = \operatorname*{arg\,min}_{P_{W|S}} \left(\mathbb{E}_{P_{W,S}}[L_E(W,S)] + \frac{1}{\gamma}\mathrm{I}(S;W) \right)$$

  - inverse temperature $\gamma \geq 0$ balances between fitting and generalization
- Replacing $\mathrm{I}(S;W)$ with $\mathrm{KL}(P_{W|S}\|\pi(W)|P_S)$ for any prior $\pi(W)$
- It gives **information risk minimization** (IRM) problem

$$P^{\star}_{W|S} = \operatorname*{arg\,min}_{P_{W|S}} \left(\mathbb{E}_{P_{W,S}}[L_E(W,S)] + \frac{1}{\gamma}\mathrm{KL}(P_{W|S}\|\pi(W)|P_S) \right)$$

Information Risk Minimization

Lemma ([Zhang, 2006, Xu and Raginsky, 2017])

Solution to IRM problem is $(\gamma, \pi(w), L_E(w,s))$-**Gibbs distribution**

$$P^{\gamma}_{W|S}(w|s) \triangleq \frac{\pi(w)e^{-\gamma L_E(w,s)}}{V(s,\gamma)}, \quad \gamma \geq 0,$$

where $V(s,\gamma) \triangleq \int \pi(w)e^{-\gamma L_E(w,s)}dw$ *is partition function.*

Proof.

For any learning algorithm $P_{W|S}$ with fixed $S = s$,

$$\begin{aligned} 0 &\leq \mathrm{KL}(P_{W|S=s} \| P^{\gamma}_{W|S=s}) \\ &= \mathbb{E}_{P_{W|S=s}}\left[\log \frac{P_{W|S=s} \cdot V(s,\gamma)}{\pi(W) \cdot e^{-\gamma L_E(w,s)}}\right] \\ &= \mathrm{KL}(P_{W|S=s} \| \pi(W)) + \log V(s,\gamma) + \gamma \mathbb{E}_{P_{W|S=s}}[L_E(w,s)]. \end{aligned}$$

$$\min_{P_{W|S}} \mathbb{E}_{P_{W|S=s}}[L_E(W,s)] + \frac{1}{\gamma}\mathrm{KL}(P_{W|S=s} \| \pi) = -\frac{1}{\gamma}\log V(s,\gamma). \quad \square$$

Gibbs Algorithm

We focus on the generalization error of Gibbs algorithm (distribution)

$(\gamma, \pi(w), L_E(w, s))$-Gibbs distribution:

$$P^{\gamma}_{W|S}(w|s) \triangleq \frac{\pi(w)e^{-\gamma L_E(W,s)}}{V(s,\gamma)}, \quad \gamma \geq 0$$

where

- inverse temperature γ, reduces to standard ERM if $\gamma \to \infty$
- $\pi(w)$ arbitrary prior distribution of W
- $V(s,\gamma) \triangleq \int \pi(w)e^{-\gamma L_E(w,s)} dw$ partition function

Practical Implementation of Gibbs algorithm

- Stochastic Gradient Langevin Dynamics (SGLD)
- Metropolis adjusted Langevin algorithm (MALA)

The SGLD can be viewed as the noisy version of SGD,

$$W_{k+1} = W_k - \eta_t \nabla L_E(W_k, s) + \sqrt{\frac{2\eta_t}{\gamma}}\zeta_k, \quad k = 0, 1, \cdots,$$

where ζ_k standard Gaussian random vector; $\eta_t > 0$ step size.

- [Raginsky et al., 2017] shows that $P_{W_k|S}$ induced by SGLD converges to $(\gamma, \pi(W_0), L_E(w_k, s))$-Gibbs distribution for sufficiently large k
- MALA is SGLD with Metropolis rejection, faster convergence

Table of Contents (Part II)

Expected Generalization Error

An **exact** characterization of generalization error for Gibbs algorithm

Theorem

For $(\gamma, \pi(w), L_E(w,s))$*-Gibbs algorithm,*

$$P^{\gamma}_{W|S}(w|s) = \frac{\pi(w) e^{-\gamma L_E(w,s)}}{V(s,\gamma)}, \quad \gamma > 0,$$

the expected generalization error is

$$\overline{\text{gen}}(P^{\gamma}_{W|S}, P_S) = \frac{\mathrm{I}_{\mathrm{SKL}}(W;S)}{\gamma}.$$

- Highlights the fundamental role of $\mathrm{I}_{\mathrm{SKL}}(W;S)$ in learning theory
- Holds even for non-i.i.d training samples

G. Aminian*, Y. Bu*, L. Toni, M. R. Rodrigues, G. W. Wornell. "An Exact Characterization of the Generalization Error for the Gibbs Algorithm," in *Proc. Conference on Neural Information Processing Systems (***NeurIPS***)*, Dec. 2021.

Generalization Error of Gibbs Algorithm

Theorem

For Gibbs algorithm $P^{\gamma}_{W|S}(w|s) = \frac{\pi(w)e^{-\gamma L_E(w,s)}}{V(s,\gamma)}$,

$$\overline{\text{gen}}(P^{\gamma}_{W|S}, P_S) = \frac{\text{I}_{\text{SKL}}(W;S)}{\gamma}.$$

Sketch of Proof:

Symmetrized KL information can be written as

$$\begin{aligned}\text{I}_{\text{SKL}}(W;S) &= \mathbb{E}_{P_{W,S}}[\log(\frac{P^{\gamma}_{W|S}}{P_W})] + \mathbb{E}_{P_W \otimes P_S}[\log(\frac{P_W}{P^{\gamma}_{W|S}})] \\ &= \mathbb{E}_{P_{W,S}}[\log(P^{\gamma}_{W|S})] - \mathbb{E}_{P_W \otimes P_S}[\log(P^{\gamma}_{W|S})]\end{aligned}$$

Note that $P_{W,S}$ and $P_W \otimes P_S$ share the same marginal distribution,

$$\begin{aligned}\text{I}_{\text{SKL}}(W;S) &= \mathbb{E}_{P_{W,S}}[-\gamma L_E(W,S)] - \mathbb{E}_{P_W \otimes P_S}[-\gamma L_E(W,S)] \\ &= \gamma \overline{\text{gen}}(P^{\gamma}_{W|S}, P_S)\end{aligned}$$

□

Empirical risk of Gibbs algorithm

Theorem

$\log V(s,\gamma)$ is convex and differentiable infinitely many times with respect to γ. In particular,

$$\mathbb{E}_\gamma[L_E(W,s)] = -\frac{\partial \log V(s,\gamma)}{\partial \gamma},$$

$$\mathrm{Var}_\gamma[L_E(W,s)] = \frac{\partial^2 \log V(s,\gamma)}{\partial \gamma^2},$$

where $\mathbb{E}_\gamma[\,\cdot\,] \triangleq \mathbb{E}_{P^\gamma_{W|S=s}}[\,\cdot\,]$, and $\mathrm{Var}_\gamma[L_E(W,s)] \triangleq \mathbb{E}_\gamma[L_E(W,s)^2] - \mathbb{E}_\gamma[L_E(W,s)]^2$.

Expected empirical risk of the Gibbs algorithm is non-increasing w.r.t γ

- Monoticity: $L_E(W,s)$ is non-increasing with γ
- Sub-Gaussianity: $L_E(W,s)$ is sub-Gaussian under Gibbs algorithm if $\mathrm{Var}_\gamma[L_E(W,s)]$ is bounded

---

Perlaza, Samir M., Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, and Stefano Rini. "Empirical risk minimization with relative entropy regularization," *IEEE Trans. Inf. Theory*, 2024.

Table of Contents (Part II)

Tighter Generalization Error Upper Bounds

Why do we care about upper bounds when we have exact characterization?

- Quantify how $\overline{\text{gen}}(P^{\gamma}_{W|S}, P_S)$ depends on number of i.i.d. samples n
- Useful when directly evaluating $\mathrm{I}_{\mathrm{SKL}}(W;S)$ is hard

Theorem

Suppose that

- $S = \{Z_i\}_{i=1}^{n}$ *are i.i.d generated from the distribution* P_Z
- $\ell(w, Z)$ *is* σ*-sub-Gaussian*
- $C_E \leq \frac{\mathrm{L}(W;S)}{\mathrm{I}(W;S)}$ *for some* $C_E \geq 0$*,*

$$\overline{\text{gen}}(P^{\gamma}_{W|S}, P_S) \leq \frac{2\sigma^2\gamma}{(1 + C_E)n}.$$

G. Aminian*, Y. Bu*, L. Toni, M. R. Rodrigues, G. W. Wornell. “An Exact Characterization of the Generalization Error for the Gibbs Algorithm,” in *Proc. Conference on Neural Information Processing Systems (* **NeurIPS** *)*, Dec. 2021.

Tighter Generalization Error Upper Bounds

Sketch of Proof:

Recall the mutual information-based bound,

$$\sqrt{\frac{2\sigma^2}{n}\mathrm{I}(S;W)} \geq \overline{\mathrm{gen}}(P^{\gamma}_{W|S}, P_S) = \frac{\mathrm{I}(W;S) + L(W;S)}{\gamma}$$

$$\geq \frac{(1+C_E)}{\gamma}\mathrm{I}(W;S)$$

$$\overline{\mathrm{gen}}(P^{\gamma}_{W|S}, P_S) \leq \sqrt{\frac{2\sigma^2}{n}\mathrm{I}(S;W)} \leq \frac{2\sigma^2\gamma}{(1+C_E)n} \qquad \square$$

[Choice of C_E]

- $C_E = 0$ is always valid, which gives $\overline{\mathrm{gen}}(P^{\gamma}_{W|S}, P_S) \leq \frac{2\sigma^2\gamma}{n}$
- $C_E = 1$, $L(S;W) \geq \mathrm{I}(S;W)$ holds for any Gaussian channel $P_{W|S}$

Example: Mean Estimation

- Learning mean $\boldsymbol{\mu} \in \mathbb{R}^d$ of Z using n i.i.d training samples $S = \{\boldsymbol{z}_i\}_{i=1}^n$
- Not necessary Gaussian, but covariance matrix $\Sigma_Z = \sigma_Z^2 \boldsymbol{I}_d$
- Mean-squared loss $\ell(\boldsymbol{w}, \boldsymbol{z}) = \|\boldsymbol{z} - \boldsymbol{w}\|_2^2$
- Gaussian prior $\pi(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 \boldsymbol{I}_d)$
- Then, $(\gamma, \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 \boldsymbol{I}_d), L_E(\boldsymbol{w}, s))$-Gibbs algorithm is given by the following Gaussian posterior

$$P_{W|S}^{\gamma}(\boldsymbol{w}|\boldsymbol{z}^n) \sim \mathcal{N}\Big(\alpha \boldsymbol{\mu}_0 + (1-\alpha)\bar{\boldsymbol{z}}, \alpha \sigma_0^2 \boldsymbol{I}_d\Big),$$

with

$$\alpha \triangleq \frac{1}{2\sigma_0^2 \gamma + 1}, \quad \bar{\boldsymbol{z}} \triangleq \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i.$$

Example: Mean Estimation

Since $P^{\gamma}_{W|S}$ is Gaussian,

$$\mathrm{I}(S;W) = \frac{d\sigma_0^2\sigma_Z^2\gamma}{(n\sigma_0^2 + \frac{1}{2\gamma})} - \mathrm{KL}\big(P_W \| \mathcal{N}(\boldsymbol{\mu}_W, \sigma_1^2 I_d)\big),$$
$$L(S;W) = \frac{d\sigma_0^2\sigma_Z^2\gamma}{(n\sigma_0^2 + \frac{1}{2\gamma})} + \mathrm{KL}\big(P_W \| \mathcal{N}(\boldsymbol{\mu}_W, \sigma_1^2 I_d)\big),$$

with $\boldsymbol{\mu}_W = \alpha\boldsymbol{\mu}_0 + (1-\alpha)\boldsymbol{\mu}$.

The generalization error can be computed exactly as:

$$\overline{\mathrm{gen}}(P^{\gamma}_{W|S}, P_S) = \frac{\mathrm{I}_{\mathrm{SKL}}(W;S)}{\gamma} = \frac{2d\sigma_0^2\sigma_Z^2}{n(\sigma_0^2 + \frac{1}{2\gamma})}.$$

As a comparison, the ISMI-based bound gives a sub-optimal bound $\mathcal{O}\left(1/\sqrt{n}\right)$, as $n \to \infty$.

Check Point

Generalization error or empirical risk is one part of the story

Our goal is to **design** (or guide the design) algorithms that minimize population risk.

There are three elements in $(\gamma, \pi(w), L_E(w, s))$-Gibbs algorithm

- inverse temperature $\gamma \longrightarrow$ Optimal hyper-parameter
- empirical risk $L_E(w, s)$, or model family $\longrightarrow$ Information criteria for model selection
- prior distribution $\pi(w) \longrightarrow$ Transfer learning

Table of Contents (Part II)

Expected Test Loss

For fixed training data s and testing data s', consider expected test loss

$$L_P(\gamma, s, s') \triangleq \mathbb{E}_\gamma[L_E(W, s')],$$

and expected generalization error

$$\overline{\mathrm{gen}}(\gamma, s, s') \triangleq \mathbb{E}_\gamma[L_E(W, s') - L_E(W, s)].$$

Theorem

For $\gamma \geq 0$ such that $\log V(s, \gamma) < \infty$, the first order derivative of the expected test loss is given by

$$\frac{\partial}{\partial \gamma} L_P(\gamma, s, s') = -\mathrm{Cov}_\gamma[L_E(W, s'), L_E(W, s)],$$

with

$$\mathrm{Cov}_\gamma[L_E(W, s'), L_E(W, s)] \triangleq \mathbb{E}_\gamma[L_E(W, s) L_E(W, s')] - \mathbb{E}_\gamma[L_E(W, s)]\mathbb{E}_\gamma[L_E(W, s')].$$

$\mathrm{Cov}_\gamma[L_E(W, s'), L_E(W, s)]$ can be positive/negative, no monotonicity

---

Y. Bu, "Towards Optimal Inverse Temperature in the Gibbs Algorithm," in IEEE ISIT 2024

Expected Generalization Error

Corollary

For $\gamma \geq 0$ such that $\log V(s,\gamma) < \infty$, the first order derivative of the expected generalization error is given by

$$\frac{\partial}{\partial\gamma}\overline{\mathrm{gen}}(\gamma, s, s') = \mathrm{Var}_\gamma(L_E(W,s)) - \mathrm{Cov}_\gamma[L_E(W,s'), L_E(W,s)].$$

- Cannot show that the $\overline{\mathrm{gen}}$ is non-decreasing, Cauchy-Schwarz Inequality only guarantees that

$$\left|\mathrm{Cov}_\gamma[L_E(W,s'), L_E(W,s)]\right| \leq \sqrt{\mathrm{Var}_\gamma(L_E(W,s))\mathrm{Var}_\gamma(L_E(W,s'))}.$$

- [Aminian et al., 2021] provides a bound of order $\mathcal{O}\left(\frac{\gamma}{n}\right)$ by simply combining the I_{SKL} characterization with the MI bound, which may hint that $\overline{\mathrm{gen}}$ is always increasing with γ.
- However, we will illustrate how $\overline{\mathrm{gen}}$ rises from zero and then decreases as γ increases.

Example: Mean Estimation

$(\gamma, \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 \boldsymbol{I}_d), L_E(\boldsymbol{w}, s))$-Gibbs algorithm is given by the following Gaussian posterior

$$P^{\gamma}_{W|S}(\boldsymbol{w}|\boldsymbol{z}^n) \sim \mathcal{N}\Big(\alpha \boldsymbol{\mu}_0 + (1-\alpha)\bar{\boldsymbol{z}}, \alpha\sigma_0^2 \boldsymbol{I}_d\Big)$$

Population risk has the following exact characterization

$$L_P(P^{\gamma}_{W|S}, P_S) = \underbrace{\frac{4d\sigma_0^2\sigma_Z^2\gamma}{n(1+2\sigma_0^2\gamma)}}_{\text{generalization error}} + \underbrace{\frac{\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}\|_2^2 + d\sigma_z^2/n}{(1+2\sigma_0^2\gamma)^2} + \frac{d\sigma_0^2}{1+2\sigma_0^2\gamma} + \frac{n-1}{n} d\sigma_Z^2.}_{\text{empirical risk}}$$

To find optimal γ minimizes L_P

- Optimize over γ using the above equation directly
- Evaluate the derivative of $L_P(\gamma, s, s')$ by computing covariance

Example: Mean Estimation

γ^* depends on other parameters of the problem in a non-trivial manner

$$\gamma^* = \begin{cases} +\infty, & \text{if } \frac{\sigma_Z^2}{n} \in [0, \frac{\sigma_0^2}{2}), \text{ (high-SNR)} \\ \frac{\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|^2 + d\sigma_0^2/2}{d(2\sigma_Z^2/n - \sigma_0^2)\sigma_0^2}, & \text{if } \frac{\sigma_Z^2}{n} \in [\frac{\sigma_0^2}{2}, \infty). \text{ (low-SNR)} \end{cases}$$

- $\frac{\sigma_Z^2}{n}$ only depends on S, can be interpreted as normalized noise
- σ_0^2 and $\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2$ captures the confidence and bias of prior knowledge
- high-SNR regime, high-quality training samples, discarding prior distribution and employing standard ERM
- low-SNR regime, where we should incorporate knowledge from both training samples and prior, optimal γ depends on everything
- If $\boldsymbol{\mu}_0 = \boldsymbol{\mu}$ and $\sigma_0^2 = 0$, $\gamma^* = 0$

Example: Linear Regression

- Training data $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, with $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$
- Data is generated using true weights $W^* \in \mathbb{R}^d$ with additive noise,

$$Y_i = X_i \cdot W^* + \varepsilon_i, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2).$$

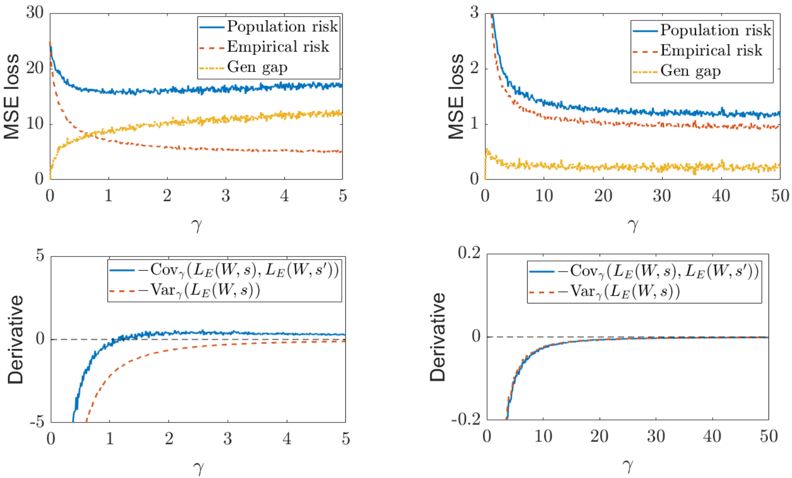
- Mean-squared loss $\ell(\boldsymbol{w}, \boldsymbol{z}) = (y - \boldsymbol{x} \cdot \boldsymbol{w})^2$
- Gaussian prior $\pi(\boldsymbol{w}) = \mathcal{N}(0, \sigma_0^2 \boldsymbol{I}_d)$
- $(\gamma, \mathcal{N}(0, \sigma_0^2 \boldsymbol{I}_d), L_E(\boldsymbol{w}, s))$-Gibbs algorithm is Gaussian

$$P_{W|S}^{\gamma}(\boldsymbol{w}|S) \sim \mathcal{N}\left(\boldsymbol{\Sigma}^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \frac{n}{2\gamma} \boldsymbol{\Sigma}^{-1}\right),$$

with $\boldsymbol{\Sigma} \triangleq \boldsymbol{X}^\top \boldsymbol{X} + \frac{n}{2\sigma_0^2 \gamma} \boldsymbol{I}_d$, and $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{Y} \in \mathbb{R}^n$ are the matrix form of the training data.

Simulation of Linear Regression

Low SNR regime, $n = 10$ and $\sigma_\varepsilon^2 = 3$; high SNR regime, $n = 100$ and $\sigma_\varepsilon^2 = 1$.



Low-SNR High-SNR

Table of Contents (Part II)

Asymptotic Behavior of Generalization Error

- Can we show something for ERM by letting $\gamma \to \infty$?
  - Previous upper bound has order $\mathcal{O}(\frac{\gamma}{n})$
- Asymptotic normality of Gibbs algorithm
  - **Single-well** case: there exists a unique $W^*(S)$

$$W^*(S) = \arg\min_{w \in \mathcal{W}} L_E(w, S).$$

  - If $H^*(S) \triangleq \nabla_w^2 L_E(w, S)\big|_{w=W^*(S)}$ is invertible [Hwang, 1980],

$$P_{W|S}^{\gamma} \to \mathcal{N}(W^*(S), \frac{1}{\gamma} H^*(S)^{-1})$$

G. Aminian*, Y. Bu*, L. Toni, M. R. Rodrigues, G. W. Wornell. "An Exact Characterization of the Generalization Error for the Gibbs Algorithm," in *Proc. Conference on Neural Information Processing Systems (*__NeurIPS__*)*, Dec. 2021.

Asymptotic Behavior of MLE

Maximum likelihood estimates (MLE) in the asymptotic regime $n \to \infty$.

- n i.i.d. training samples generated from distribution P_Z
- Fit training data with distribution family $f(z_i|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^p$
- $P_Z = f(\cdot|\boldsymbol{\theta}^*)$ for $\boldsymbol{\theta}^* \in \mathcal{W}$
- log-loss $\ell(\boldsymbol{w}, z) = -\log f(z|\boldsymbol{w})$

As $\gamma \to \infty$, Gibbs algorithm converges to ERM algorithm (MLE),

$$\hat{W}_{\mathrm{ML}} \triangleq \arg\max_{\boldsymbol{\theta} \in \mathcal{W}} \sum_{i=1}^{n} \log f(Z_i|\boldsymbol{\theta}).$$

Compute $\mathrm{I}_{\mathrm{SKL}}(W; S)$ using Gaussian approximation

$$\overline{\mathrm{gen}}(P_{W|S}^{\infty}, P_S) = \frac{d}{n}.$$

Connection to Model Selection

- K candidate models $M_1, M_2, \ldots, M_K$
- Each model M_k is characterized by parametric probabilistic model $P_k(\boldsymbol{z}|\boldsymbol{\theta}_k)$ and prior $\pi_k(\boldsymbol{\theta}_k)$
- log likelihood as the loss function $\ell_{\log}(w, \boldsymbol{z}) \triangleq -\log P(\boldsymbol{z}|w)$

How to select the *optimal* model?

- Information Criteria for Model Selection
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)

Akaike Information Criterion (AIC)

AIC selects the model that minimizes population risk:

$$\arg\min_k \ \mathrm{KL}(P_Z \| P_k(\boldsymbol{z}|\hat{\boldsymbol{\theta}}^{(k)}_{\mathrm{ML}})) = \arg\min_k \mathbb{E}_{P_Z}\big[-\log P_k(Z|\hat{\boldsymbol{\theta}}^{(k)}_{\mathrm{ML}})\big].$$

AIC approximates it using empirical risk and generalization error

$$\mathrm{AIC} = \arg\min_k L_E(\hat{\boldsymbol{\theta}}^{(k)}_{\mathrm{ML}}, S) + \overline{gen}(\hat{\boldsymbol{\theta}}^{(k)}_{\mathrm{ML}}, P_Z).$$

In classic regime where $n \to \infty$, and certain regularization conditions

$$\mathrm{AIC} = \arg\min_k L_E(\hat{\boldsymbol{\theta}}^{(k)}_{\mathrm{ML}}, S) + \frac{p}{n}.$$

Bayesian Information Criterion (BIC)

BIC selects the model that maximizes marginal likelihood:

$$m_k(\boldsymbol{z}^n) \triangleq \int P_k(\boldsymbol{z}^n|\boldsymbol{\theta}_k)\,\pi_k(\boldsymbol{\theta}_k)\,d\boldsymbol{\theta}_k,$$

which is equivalent to maximizing posterior probability $P(M_k|\boldsymbol{z}^n)$.

$$\begin{aligned}\mathrm{BIC} &= \arg\min_k \; -\frac{1}{n}\log m_k(\boldsymbol{z}^n) \\ &= \arg\min_k \; L_E(\hat{\boldsymbol{\theta}}^{(k)}_{\mathrm{ML}}, S) + \frac{p_k \log n}{2n},\end{aligned}$$

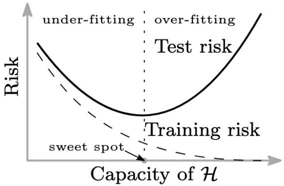
where Laplace approximation is applied as $n \to \infty$.

Comparison between AIC and BIC

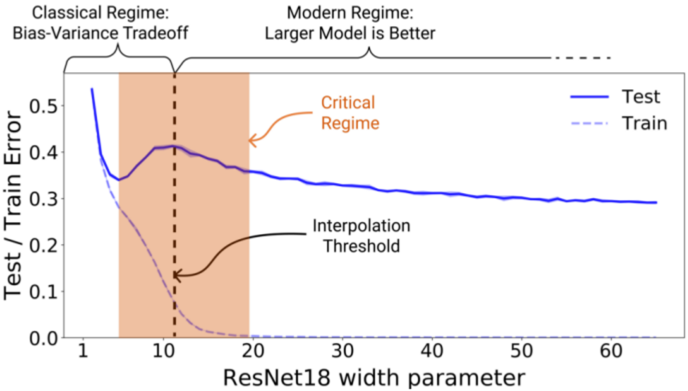
$$\mathrm{AIC} = \arg\min \; L_E(\hat{\theta}_{\mathrm{ML}}, S) + \frac{p}{n}$$

$$\mathrm{BIC} = \arg\min \; L_E(\hat{\theta}_{\mathrm{ML}}, S) + \frac{p \log n}{2n}.$$

- AIC minimizes population risk (optimal prediction performance)
- BIC maximizes the marginal likelihood (identifying the true model)
- BIC imposing a **larger** penalty for more complex models.



Double-descent in Over-parameterized Regime



- When $p \leq n$, the classical $\cup$-shaped curve is valid.
- When $p \geq n$, test loss can decrease again.

Challenges in Over-parameterized regime

Asymptotic normality (AIC) and Laplace Approximation (BIC) do not hold in this new regime!

There are some efforts to extend these information criteria:

- Akaike's Information Corrected Criterion (AICC), fixed p, small n
- Widely applicable BIC (WBIC), singular Hessian matrix

More recent work trying to demystify double-descent

- Neural Tangent Kernel (NTK), lazy training
- Random feature model
- Mean-field approach

Marginal likelihood of Gibbs algorithm

Recall the **information risk minimization** for motivating the Gibbs algorithm.

$$\min_{P_{W|S}} \mathbb{E}_{P_{W|S=s}}\left[L_E(W,s)\right] + \frac{1}{\gamma}\mathrm{KL}(P_{W|S=s}\|\pi) = -\frac{1}{\gamma}\log V(s,\gamma).$$

If we adopt log-loss function $\ell(w,\boldsymbol{z}) = -\log P(\boldsymbol{z}|w)$, and set $\gamma = n$

$$\begin{aligned}
-\frac{1}{\gamma}\log V(s,\gamma) &= -\frac{1}{n}\log\int \pi(w)e^{-nL_E(w,s)}dw \\
&= -\frac{1}{n}\log\int \pi(w)P(\boldsymbol{z}^n|w)dw \\
&= -\frac{1}{n}\log m(\boldsymbol{z}^n)
\end{aligned}$$

Gibbs based Information Criteria

Gibbs-based AIC:

$$\mathrm{AIC}^{+} \triangleq L_E(\hat{W}_{\mathrm{Gibbs}}, \boldsymbol{z}^n) + \frac{1}{n}\mathrm{I}_{\mathrm{SKL}}(P^*_{\hat{W}|S}, P_S).$$

Gibbs-based BIC:

$$\mathrm{BIC}^{+} \triangleq L_E(\hat{W}_{\mathrm{Gibbs}}, \boldsymbol{z}^n) + \frac{1}{n}\mathrm{KL}(P^*_{W|S=\boldsymbol{z}^n} \| \pi),$$

$$\mathrm{BIC}^{-} \triangleq \mathbb{E}_{\pi}\left[L_E(W, \boldsymbol{z}^n)\right] - \frac{1}{n}\mathrm{KL}(\pi \| P^*_{W|S=\boldsymbol{z}^n}).$$

We can show that in the classic regime where p is fixed and $n \to \infty$, they all reduce back to their classical forms.

H. Chen, Y. Bu, G. W. Wornell, "Gibbs-Based Information Criteria and the Over-Parameterized Regime," in *Proc. International Conference on Artificial Intelligence and Statistics (**AISTATS**)*, 2024.

Random Feature Model

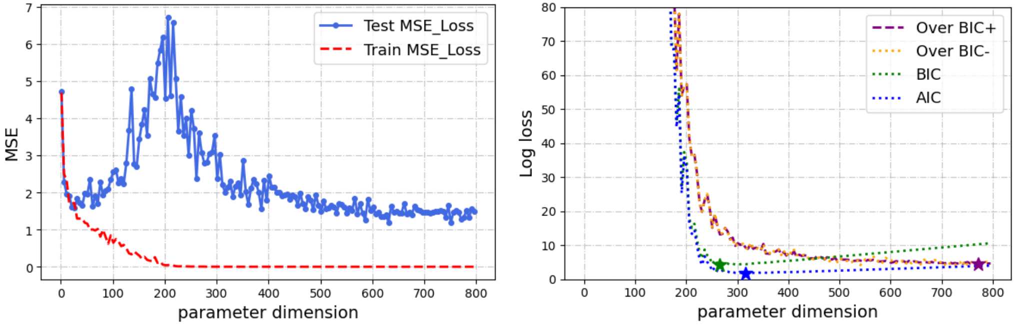
The output of **Random Feature (RF) model** with input data $\boldsymbol{x} \in \mathbb{R}^d$ is

$$g(\boldsymbol{x}) \triangleq \sum_{j=1}^{p} f\left(\frac{\langle \boldsymbol{x}, \boldsymbol{F}_j \rangle}{\sqrt{d}}\right) \boldsymbol{w}_j = f\left(\frac{\boldsymbol{x}^\top \boldsymbol{F}}{\sqrt{d}}\right) \boldsymbol{w},$$

- Two-layer neural network with i.i.d Gaussian weights $\boldsymbol{F} \in \mathbb{R}^{d \times p}$ in the first layer, only the second layer is trainable
- $f()$ is the non-linear activation function
- The dimensionality of input data d is not entangled with number of parameters p

Experiment

Evaluating the BIC^{+} and BIC^{-} using $n = 200$ samples in RF models



- We observe **Double-descent** in population risk for RF model
- Our Gibbs-based BICs prefer **over-parameterized models**

Check point

- Provide information criteria for the Gibbs algorithm, with different **information measures** as the penalty terms.

- Generalize our information-theoretic analysis to **over-parameterized** random feature.

- The **mismatch** between **marginal likelihood (BIC)** and **generalization error (AIC)** in the over-parameterized setting, which highly depends on the prior distributions.

Table of Contents (Part II)

Generalization of Transfer Learning

- Source data set $D^s = \{Z_i^s\}_{i=1}^m$, generated from P_{D^s}
- Target data set $D^t = \{Z_j^t\}_{j=1}^n$, generated from P_{D^t}
- The empirical risk of source and target task

$$L_E(w, d^s) \triangleq \frac{1}{m}\sum_{j=1}^{m} \ell(w, z_j^s), \qquad L_E(w, d^t) \triangleq \frac{1}{n}\sum_{j=1}^{n} \ell(w, z_j^t).$$

- The population risk of the target task

$$L_P(w, P_{D^t}) \triangleq \mathbb{E}_{P_{D^t}}[L_E(w, D^t)].$$

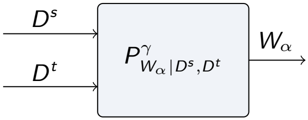
- Expected Transfer Generalization Error

$$\overline{\mathrm{gen}}(P_{W|D^s,D^t}, P_{D^s}, P_{D^t}) \triangleq \mathbb{E}_{P_{W,D^s,D^t}}[L_P(W, P_{D^t}) - L_E(W, D^t)].$$

Transfer Learning: α-Weighted ERM

- Output hypothesis w_α is trained by minimizing a convex combination of the source and target task empirical risks [Ben-David et al., 2010], for $\alpha \in [0,1]$

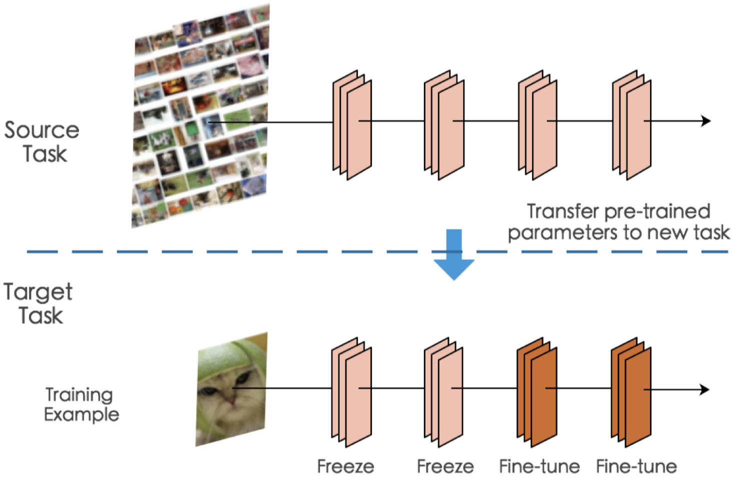
$$L_E(w_\alpha, d^s, d^t) = (1-\alpha)L_E(w_\alpha, d^s) + \alpha L_E(w_\alpha, d^t)$$



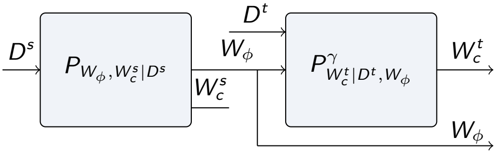
- **α-weighted Gibbs algorithm** generalizes the α-weighted-ERM by considering the $(\gamma, \pi(w_\alpha), L_E(w_\alpha, d^s, d^t))$-Gibbs algorithm

$$P^{\gamma}_{W_\alpha|D^s,D^t}(w_\alpha|d^s, d^t) = \frac{\pi(w_\alpha)e^{-\gamma L_E(w_\alpha, d^s, d^t)}}{V_\alpha(d^s, d^t, \gamma)}.$$

Transfer Learning: Two-stage ERM



Two-stage-ERM Transfer Learning



- **First Stage:** Learn shared feature extractor $w_\phi \in \mathcal{W}_\phi$

$$[W_\phi, W_c^s] = \arg\min_{w} L_E^{S1}(w, d^s).$$

- **Second Stage:** Freeze W_ϕ, and learn target-specific hypothesis w_c^t

$$W_c^t = \arg\min_{w_c} L_E^{S2}([W_\phi, w_c], d^t)$$

Expected Transfer Generalization Error

Theorem

The expected transfer generalization error of the α-weighted Gibbs algorithm is given by

$$\overline{\text{gen}}_\alpha(P_{D^s}, P_{D^t}) = \frac{\mathrm{I}_{\mathrm{SKL}}(W_\alpha; D^t|D^s)}{\alpha\gamma}.$$

Theorem

The expected transfer generalization error of the two-stage Gibbs algorithm is given by

$$\overline{\text{gen}}_\beta(P_{D^s}, P_{D^t}) = \frac{\mathrm{I}_{\mathrm{SKL}}(W_c^t; D^t|W_\phi)}{\gamma}.$$

Y. Bu*, G. Aminian*, L. Toni, M. R. Rodrigues, G. W. Wornell. "Characterizing and Understanding the Generalization Error of Transfer Learning with Gibbs Algorithm," in *Proc. International Conference on Artificial Intelligence and Statistics (***AISTATS***)* 2022.

Asymptotic Behavior of MLE

Maximum likelihood estimates

- n i.i.d. target samples, m i.i.d. source samples
- Fit training data with distribution family $f(z|\boldsymbol{w})$, $\boldsymbol{w} = (\boldsymbol{w}_\phi, \boldsymbol{w}_c) \in \mathbb{R}^d$, $\boldsymbol{w}_c \in \mathbb{R}^{d_c}$
- $P_{Z^t} = f(\cdot|\boldsymbol{w}^*)$ for $\boldsymbol{w}^* \in \mathcal{W}$
- log-loss $\ell(\boldsymbol{w}, z) = -\log f(z|\boldsymbol{w})$

	Standard target ERM	α-weighted ERM	Two-stage ERM
$\overline{\mathrm{gen}}$	$\mathcal{O}(\frac{d}{n})$	$\mathcal{O}(\frac{d}{m+n})$	$\mathcal{O}(\frac{d_c}{n})$

Table of Contents (Part II)

Conclusion

- Connect **operational quantity** in learning theory (generalization error, marginal likelihood) with different **information measures** for Gibbs algorithm
- Demonstrate the **versatility** of our approach in multiple applications
  - Optimal Inverse temperature
  - Gibbs-based BIC for over-parameterized model selection
  - Gibbs based-**transfer** learning
- Our Gibbs-based analysis provides an information-theoretic **framework** for understanding generalization behavior in modern machine learning, still a lot to be explored!

Q & A

Thank you for your attention!

References I

Aminian, G., Arjmandi, H., Gohari, A., Nasiri-Kenari, M., and Mitra, U. (2015).
Capacity of diffusion-based molecular communication networks over lti-poisson channels.
IEEE Transactions on Molecular, Biological and Multi-Scale Communications, 1(2):188–201.

Aminian, G., Bu, Y., Toni, L., Rodrigues, M., and Wornell, G. (2021).
An exact characterization of the generalization error for the gibbs algorithm.
Advances in Neural Information Processing Systems, 34:8106–8118.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010).
A theory of learning from different domains.
Machine learning, 79(1):151–175.

References II

Bu, Y., Zou, S., and Veeravalli, V. V. (2020).

Tightening mutual information-based bounds on generalization error.

IEEE Journal on Selected Areas in Information Theory, 1(1):121–130.

Harutyunyan, H., Raginsky, M., Ver Steeg, G., and Galstyan, A. (2021).

Information-theoretic generalization bounds for black-box learning algorithms.

Advances in Neural Information Processing Systems, 34:24670–24682.

Hwang, C.-R. (1980).

Laplace's method revisited: weak convergence of probability measures.

The Annals of Probability, pages 1177–1182.

References III

Palomar, D. P. and Verdú, S. (2008).

Lautum information.

IEEE transactions on information theory, 54(3):964–975.

Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017).

Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis.

In *Conference on Learning Theory*, pages 1674–1703. PMLR.

Steinke, T. and Zakynthinou, L. (2020).

Reasoning about generalization via conditional mutual information.

arXiv preprint arXiv:2001.09122.

References IV

Wang, Z. and Mao, Y. (2023).

Tighter information-theoretic generalization bounds from supersamples.

In *International Conference on Machine Learning*, pages 36111–36137. PMLR.

Xu, A. and Raginsky, M. (2017).

Information-theoretic analysis of generalization capability of learning algorithms.

In *Advances in Neural Information Processing Systems*, pages 2524–2533.

Zhang, T. (2006).

Information-theoretic upper and lower bounds for statistical estimation.

IEEE Transactions on Information Theory, 52(4):1307–1321.