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 20th of November, 2024
Shenzhen, China

Gholamali Aminian

gaminian@turing.ac.uk

The Alan Turing Institute

London, UK

Aminian's research is funded by The Alan Turing Institute.

Yuheng Bu

buyuheng@ufl.edu

Department of Electrical & Computer Engineering

University of Florida.

Iñaki Esnaola

esnaola@sheffield.ac.uk

School of Electrical and Electronic Engineering

University of Sheffield

Samir M. Perlaza

samir.perlaza@inria.fr

Centre Inria d'Université Côte d'Azur

INRIA, Sophia Antipolis, France

Perlaza's research is funded in part by the **European Commission** under grant 872172; in part by the **French National Agency for Research (ANR)** under grants n°ANR-22-PEFT-0010 and ANR-21-CE25-0013; and in part by the **Agence de l'Innovation de Défense (AID)**.

Organization of the Tutorial

- **8:30-9:00** Part I: Generalization Error in Supervised Learning
- **9:05-10:00** Part II: The Gibbs Algorithm and its Generalization Error
- **10:00-10:20** Break
- **10:20-11:00** Part III: Empirical Risk Minimization and **Generalization Error**
- **11:00-11:50** Part IV: **Generalization Error** of General Machine Learning Algorithms

Organization of the Tutorial

PART I and Part II

- **8:30-9:00** Part I: Generalization Error in Supervised Learning
  - Problem formulation
  - Classic Generalization Analysis
- **9:05-10:00** Part II: The Gibbs Algorithm and its Generalization Error
  - Exact Characterization of the Gibbs algorithm
  - The Gibbs-based Information criteria (AIC and BIC)
  - Understanding Transfer Learning via the Gibbs Algorithm
- **10:00-10:20** Break

Organization of the Tutorial

PART III and Part IV

- **10:20-11:00** Part III: Empirical Risk Minimization and Generalization Error
  - Empirical Risk Minimization with f-**Divergence Regularization**
  - The **Asymmetry** of Relative Entropy
  - **Equivalences** among Empirical Risk Minimizations with different f-Divergence Regularizations
- **11:00-11:50** Part IV: Generalization Error of Machine Learning Algorithms
  - Empirical Risk **Optimization** with Relative Entropy Regularization
  - The **Method of Gaps**
  - **Equivalent Expressions** for the Generalization Error

Slides for the Tutorial

Slides for Part I

Table of Contents (Part I)

Generalization Error in Supervised Learning

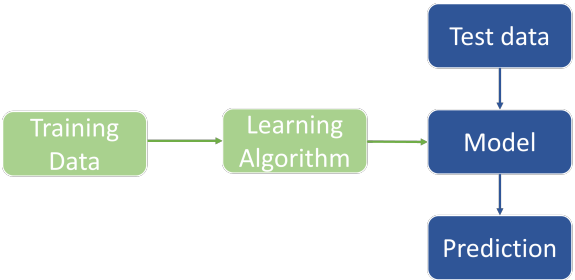
- Problem Formulation
- Different Types of Bounds

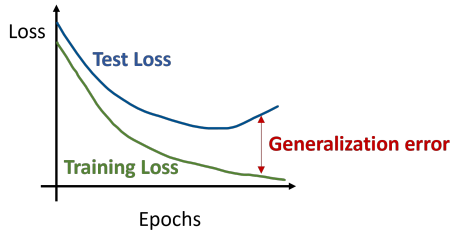
Classical Generalization Analysis

- Uniform Convergence
- Stability
- Information-theoretic Bounds

Supervised Learning

Generalization Error





Generalization error = Population risk (Test Loss) - Empirical risk (Training Loss)

Supervised Learning

Problem Formulation

- Training data set $S = \{Z_1, \cdots, Z_n\}$, $Z_i = \{X_i, Y_i\} \in \mathcal{Z}$ generated from P_S
- Parameters (weights) of learning model $w \in \mathcal{W}$, e.g., $\hat{Y} = f(X; w)$
- Nonnegative loss function $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$, e.g., $\ell(w, z) = (y - f(x; w))^2$

Empirical risk (training loss):

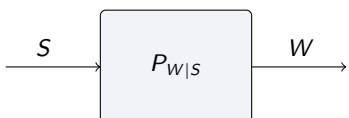
$$L_E(w, s) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i), \qquad \forall w \in \mathcal{W}$$

Population risk (test loss):

$$L_P(w, P_S) \triangleq \mathbb{E}_{P_S}[L_E(w, S)], \qquad \forall w \in \mathcal{W}$$

Learning algorithm can be modeled as randomized mapping: $P_{W|S}$.

- Randomness in initialization
- Stochastic gradient descent (SGD)
- Empirical Risk Minimization (ERM) is a special case



Generalization error:

$$\mathrm{gen}(P_{W|S}, P_S) \triangleq L_P(W, P_S) - L_E(W, S),$$

with W generated from $P_{W|S}$

Generalization Error in Supervised Learning

Different Types of Bounds

- **Single-draw Generalization Error Upper Bound:** Under joint distribution of $P_{W,S}$, following upper bound holds with probability at least $(1-\delta)$,

$$\mathrm{gen}(P_{W|S}, P_S) \leq g(\delta, n),$$

for a given real function g and $\delta \in (0,1)$,

- **PAC-Bayesian Generalization Error Upper Bound:** Under distribution P_S, following upper bound holds with probability at least $(1-\delta)$,

$$\mathbb{E}_{P_{W|S}}[\mathrm{gen}(P_{W|S}, P_S)] \leq f(\delta, n),$$

for a given real function f and $\delta \in (0,1)$,

- **Expected Generalization error Upper Bound:** The expectation of generalization error with respect to joint distribution $P_{W,S}$

$$\overline{\mathrm{gen}}(P_{W|S}, P_S) \triangleq \mathbb{E}_{P_{W,S}}[L_P(W, P_S) - L_E(W, S)] \leq h(n),$$

for a given real function h.

Table of Contents (Part I)

Generalization Error in Supervised Learning

- Problem Formulation
- Different Types of Bounds

Classical Generalization Analysis

- Uniform Convergence
- Stability
- Information-theoretic Bounds

If the **induced function** class $\mathcal{F}_{\ell,W} := \{\ell(w,\cdot) : w \in W\}$ is not 'too rich,' then

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}} |L_P(w,P_S) - L_E(w,S)|\right] \leq \frac{\mathrm{Comp}(\mathcal{F}_{\ell,W})}{\sqrt{n}},$$

where $\mathrm{Comp}(\mathcal{F}_{\ell,W})$ measures complexity of $\mathcal{F}_{\ell,W}$ and does not depend on μ (distribution-free)

Some examples:

- Cardinality of $\mathcal{F}_{\ell,W}$
- VC-dimension [Vapnik, 1999]
- Natarajan-dimension [Holden and Niranjan, 1995]
- Empirical Rademacher complexity [Bartlett and Mendelson, 2002]

We can **always** bound the generalization error as

$$\overline{\mathrm{gen}}(P_{W|S}, P_S) \leq \mathbb{E}\left[\sup_{w \in \mathcal{W}} |L_P(w, P_S) - L_E(w, S)|\right]$$

... but this bound:

- relies on restricting the complexity of the hypothesis space
- ignores the learning algorithm, $P_{W|S}$
- may be too conservative if algorithm does not explore the entire $\mathcal{W}$ due to computational budget.

Learning does not require uniform convergence

One can construct examples of (ℓ, W), where uniform convergence does not hold (the upper bound does not converge to 0 as $n \to \infty$), yet learning still takes place [Shalev-Shwartz and Ben-David, 2014].

Stability quantifies the sensitivity of algorithm $P_{W|S}$ to local modifications

- replace Z_i with Z_i' in the training data S

$$(Z_1, \cdots, Z_{i-1}, Z_i, Z_{i+1}, \cdots, Z_n) \xrightarrow{P_{W|S}} W$$
$$(Z_1, \cdots, Z_{i-1}, Z_i', Z_{i+1}, \cdots, Z_n) \xrightarrow{P_{W|S}} W^{(i)}$$

- For any learning algorithm

$$\overline{\mathrm{gen}}(P_{W|S}, P_S) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(W, Z_i') - \ell(W^{(i)}, Z_i')]$$

Definition ([Bousquet and Elisseeff, 2002] Uniform Stability)

$P_{W|S}$ is ε-uniformly stable if $\sup_z \mathbb{E}[\ell(W, z) - \ell(W^{(i)}, z)] \leq \varepsilon$.

The stability of learning algorithm $P_{W|S}$ leads to generalization.

- Population risk is the expectation of $\ell(w, s)$ under product of the marginal distributions $P_W P_S$
- Empirical risk is the expectation of $\ell(w, s)$ under joint distribution $P_{W|S} P_S$

Lemma ([Xu and Raginsky, 2017])

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

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

where σ-sub-Gaussian means

$$\log \left(\mathbb{E} \left[e^{\lambda (X - \mathbb{E}(X))} \right] \right) \leq \frac{\sigma^2}{2} \lambda^2$$

- Depends on every ingredient in the supervised learning problem
- Reducing dependence between W and S leads to better generalization bound

The proof is based Donsker-Varadhan variational representation of KL divergence:

$$\mathrm{KL}(P\|Q) = \sup_{f\in\mathcal{F}} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp f(X)],$$

where $\mathcal{F}$ denotes the set of functions $f : \mathcal{X} \to \mathbb{R}$.

Proof.

- $L_E(w, S)$ is $\frac{\sigma}{\sqrt{n}}$-sub Gaussian for any fixed w.
- Set $f(w, s) = \lambda L_E(w, s) - \lambda \mathbb{E}_S[L_E(w, S)]$ in Donsker-Varadhan

Thus,

$$\begin{aligned}
\mathrm{I}(S; W) &= \mathrm{KL}(P_{W,S}\|P_W P_S) \\
&\geq \mathbb{E}_{P_{W,S}}[\lambda f(W, S)] - \log(\mathbb{E}_{P_{\bar{W}}P_{\bar{S}}} e^{\lambda f(\bar{w},\bar{s})}) \\
&\geq \lambda \mathbb{E}_{P_{W,S}}[L_E(W, S)] - \lambda \mathbb{E}[L_E(\bar{W}, \bar{S})] - \frac{\lambda^2\sigma^2}{2n}
\end{aligned}$$

This inequality holds for all $\lambda \in \mathbb{R}$, optimizing over the λ gives the final bound. □

Summary of Existing Generalization Bounds

Traditional ways of bounding generalization errors are not satisfying:

- Do not fully characterize all aspects of learning algorithm
  - only measuring complexity of **functional space** $\mathcal{W}$, e.g., VC dimension
  - only exploring properties of **learning algorithm**, e.g., uniform stability
- Information-theoretical bounds
  - depending on input distribution P_S
  - depending on learning algorithm $P_{W|S}$

  can still be loose.

Our method differs from previous generalization bounds

- instead of a loose bound for general learning algorithms
- **exact** characterization of a specific learning algorithm that has *better structure*

References I

Bartlett, P. L. and Mendelson, S. (2002).

Rademacher and gaussian complexities: Risk bounds and structural results.

Journal of Machine Learning Research, 3(Nov):463–482.

Bousquet, O. and Elisseeff, A. (2002).

Stability and generalization.

Journal of machine learning research, 2(Mar):499–526.

Holden, S. B. and Niranjan, M. (1995).

On the practical applicability of vc dimension bounds.

Neural Computation, 7(6):1265–1288.

References II

Shalev-Shwartz, S. and Ben-David, S. (2014).

Understanding machine learning: From theory to algorithms.

Cambridge university press.

Vapnik, V. N. (1999).

An overview of statistical learning theory.

IEEE transactions on neural networks, 10(5):988–999.

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.

Slides for Part II

Table of Contents (Part II)

Generalization Error of Gibbs Algorithm

- Gibbs algorithm
- Exact Characterizations
- Other Properties

Optimal Inverse Temperature

Asymptotic Behavior and Information Criteria for Model Selection

Generalization Error of Transfer learning

Conclusion

- 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)

Generalization Error of Gibbs Algorithm

- Gibbs algorithm
- Exact Characterizations
- Other Properties

Optimal Inverse Temperature

Asymptotic Behavior and Information Criteria for Model Selection

Generalization Error of Transfer learning

Conclusion

Regularized ERM problem

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

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

  - 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

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

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 $S = s$,

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

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

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)

Generalization Error of Gibbs Algorithm

- Gibbs algorithm
- Exact Characterizations
- Other Properties

Optimal Inverse Temperature

Asymptotic Behavior and Information Criteria for Model Selection

Generalization Error of Transfer learning

Conclusion

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

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

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

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

Sketch of Proof:

Symmetrized KL information can be written as

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

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

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

□

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)

Generalization Error of Gibbs Algorithm

- Gibbs algorithm
- Exact Characterizations
- Other Properties

Optimal Inverse Temperature

Asymptotic Behavior and Information Criteria for Model Selection

Generalization Error of Transfer learning

Conclusion

Tighter Generalization Error Upper Bounds

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

- Quantify how $\overline{\mathrm{gen}}(P_{W|S}^{\gamma}, P_S)$ 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*
- $C_E \leq \frac{L(W;S)}{\mathrm{I}(W;S)}$ *for some* $C_E \geq 0$,

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

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.

Sketch of Proof:

Recall the mutual information-based bound,

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

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

[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)

Generalization Error of Gibbs Algorithm

- Gibbs algorithm
- Exact Characterizations
- Other Properties

Optimal Inverse Temperature

Asymptotic Behavior and Information Criteria for Model Selection

Generalization Error of Transfer learning

Conclusion

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

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

Population risk has the following exact characterization

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

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 $S = \{(x_i, y_i)\}_{i=1}^n$, 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,

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

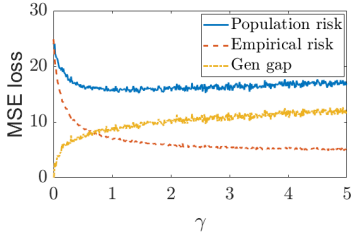
- Mean-squared loss $\ell(\boldsymbol{w}, \boldsymbol{z}) = (y - \boldsymbol{x} \cdot \boldsymbol{w})^2$
- Gaussian prior $\pi(\boldsymbol{w}) = \mathcal{N}(0, \sigma_0^2 \boldsymbol{I}_d)$
- $(\gamma, \mathcal{N}(0, \sigma_0^2 \boldsymbol{I}_d), L_E(\boldsymbol{w}, s))$-Gibbs algorithm is Gaussian

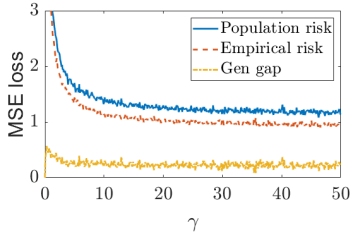
$$P_{W|S}^{\gamma}(\boldsymbol{w}|S) \sim \mathcal{N}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top \boldsymbol{Y}, \frac{n}{2\gamma}\boldsymbol{\Sigma}^{-1}\Big),$$

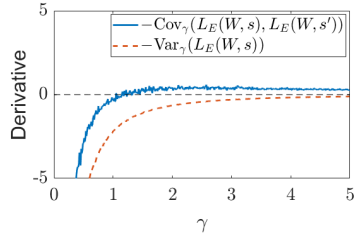
with $\boldsymbol{\Sigma} \triangleq \boldsymbol{X}^\top \boldsymbol{X} + \frac{n}{2\sigma_0^2\gamma}\boldsymbol{I}_d$, and $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{Y} \in \mathbb{R}^n$ are the matrix form of the training data.

Simulation of Linear Regression

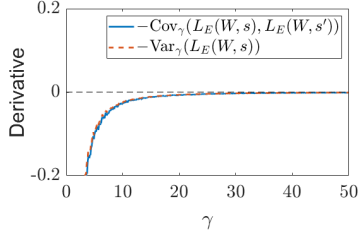
Low SNR regime, $n = 10$ and $\sigma_\varepsilon^2 = 3$; high SNR regime, $n = 100$ and $\sigma_\varepsilon^2 = 1$.







Low-SNR



High-SNR

Table of Contents (Part II)

Generalization Error of Gibbs Algorithm

- Gibbs algorithm
- Exact Characterizations
- Other Properties

Optimal Inverse Temperature

Asymptotic Behavior and Information Criteria for Model Selection

Generalization Error of Transfer learning

Conclusion

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],

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

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:

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

AIC approximates it using empirical risk and generalization error

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

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

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

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 $P(M_k|\boldsymbol{z}^n)$.

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

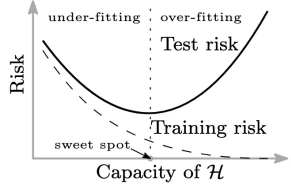
where Laplace approximation is applied as $n \to \infty$.

Comparison between AIC and BIC

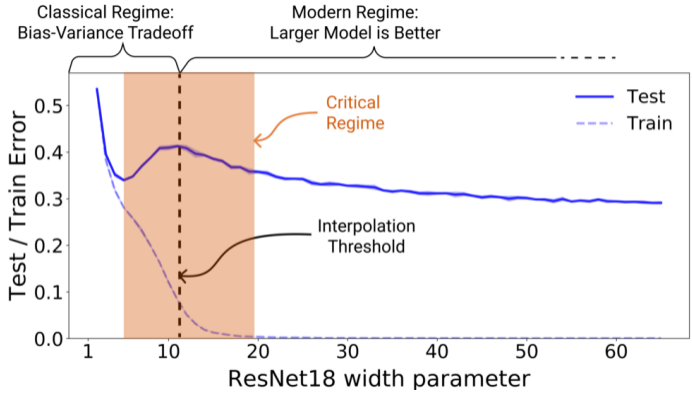
$$\text{AIC} = \arg\min \; L_E(\hat{\theta}_{\text{ML}}, S) + \frac{p}{n}$$

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

- 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.

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

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:

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

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

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

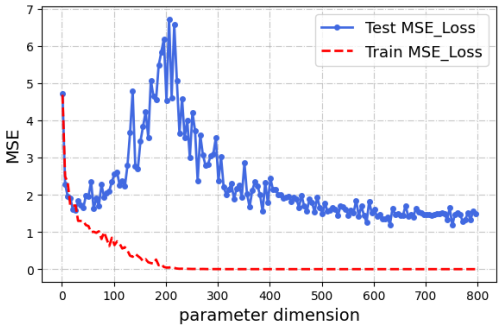
The output of **Random Feature (RF) model** with input data $\boldsymbol{x} \in \mathbb{R}^d$ is

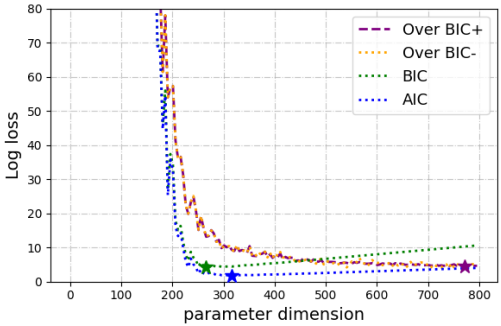
$$g(\boldsymbol{x}) \triangleq \sum_{j=1}^{p} f\Big(\frac{\langle \boldsymbol{x}, \boldsymbol{F}_j \rangle}{\sqrt{d}}\Big) \boldsymbol{w}_j = f\Big(\frac{\boldsymbol{x}^\top \boldsymbol{F}}{\sqrt{d}}\Big) \boldsymbol{w},$$

- 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 Error of Gibbs Algorithm

- Gibbs algorithm
- Exact Characterizations
- Other Properties

Optimal Inverse Temperature

Asymptotic Behavior and Information Criteria for Model Selection

Generalization Error of Transfer learning

Conclusion

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

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

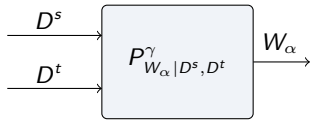
- Expected Transfer Generalization Error

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

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 $\alpha \in [0, 1]$

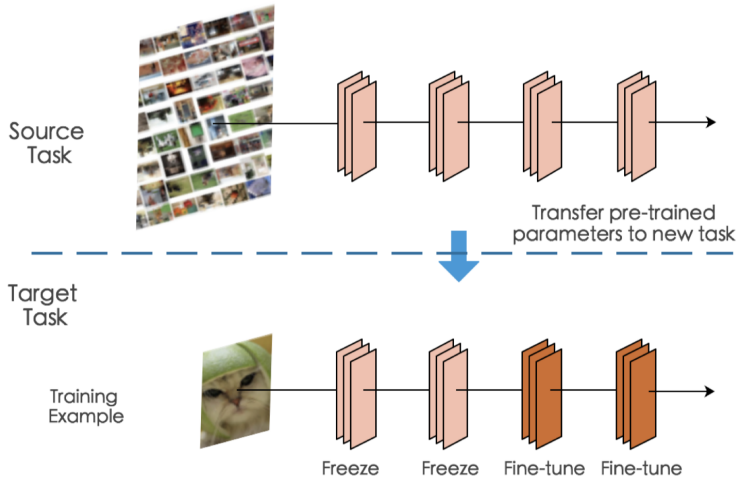
$$L_E(w_\alpha, d^s, d^t) = (1 - \alpha) L_E(w_\alpha, d^s) + \alpha L_E(w_\alpha, d^t)$$



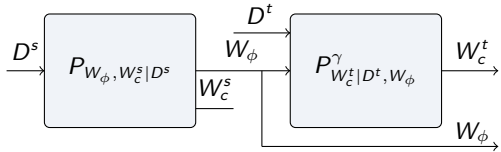
- **α-weighted Gibbs algorithm** generalizes the α-weighted-ERM by considering the $(\gamma, \pi(w_\alpha), L_E(w_\alpha, d^s, d^t))$-Gibbs algorithm

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

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

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

Theorem

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

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

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
$\overline{\text{gen}}$	$\mathcal{O}(\frac{d}{n})$	$\mathcal{O}(\frac{d}{m+n})$	$\mathcal{O}(\frac{d_c}{n})$

Table of Contents (Part II)

Generalization Error of Gibbs Algorithm

- Gibbs algorithm
- Exact Characterizations
- Other Properties

Optimal Inverse Temperature

Asymptotic Behavior and Information Criteria for Model Selection

Generalization Error of Transfer learning

Conclusion

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.

Slides for Part III

Table of Contents

Regularization in Empirical Risk Minimization

- Review of the Problem Formulation
- Empirical Risk Minimization with Relative Entropy Regularization

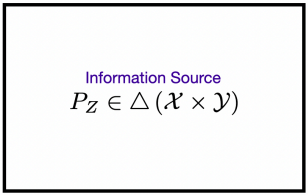
Asymmetry of Relative Entropy on the Regularization

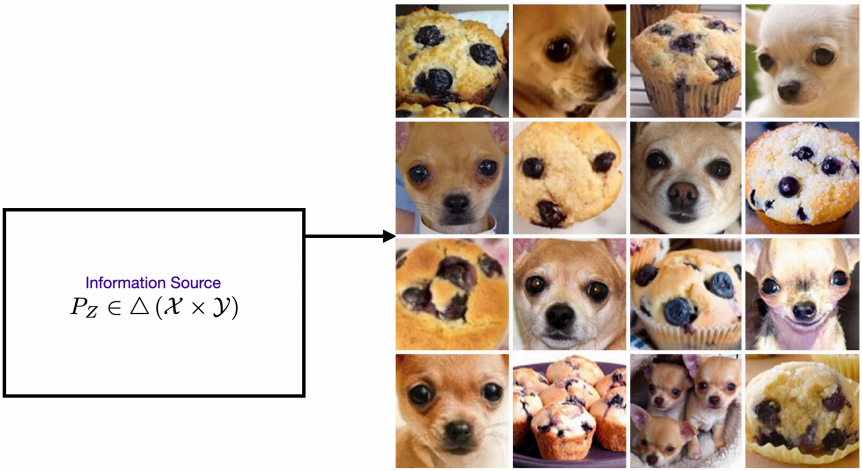
- Support Constraint of Relative Entropy Regularization
- Properties of Type-II Relative Entropy Regularization

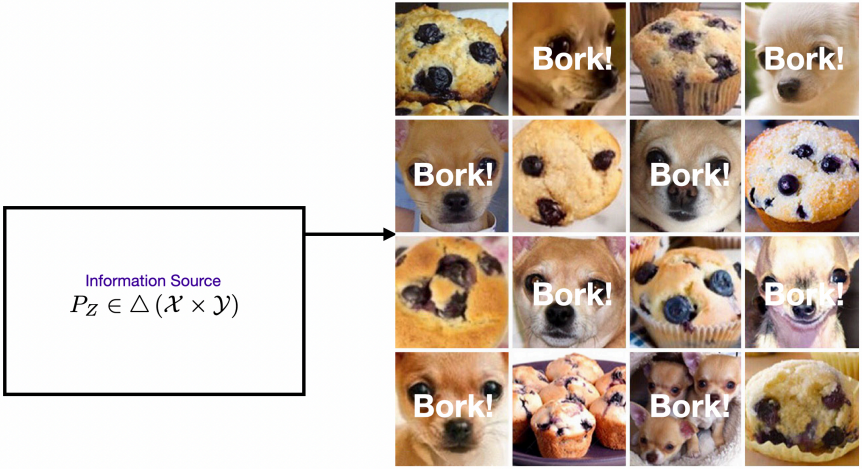
Regularization via f-divergences

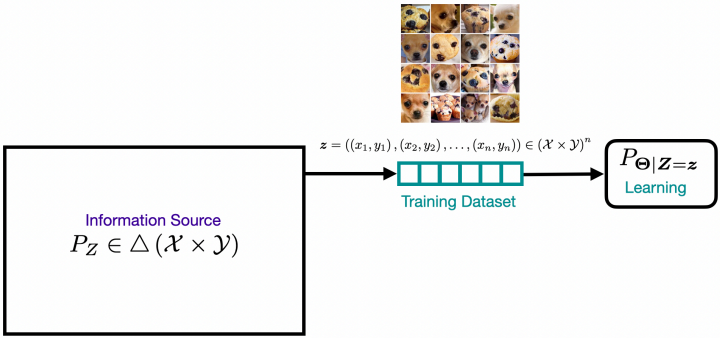
- Solution and Common Regularizers
- Equivalence of the f-Regularization via Transformation of the Empirical Risk

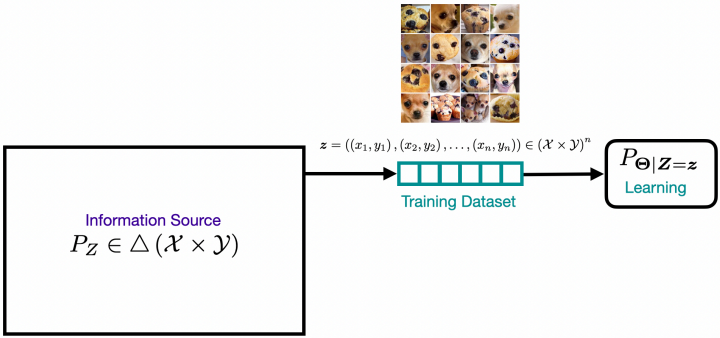
Conclusions


Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$


Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$

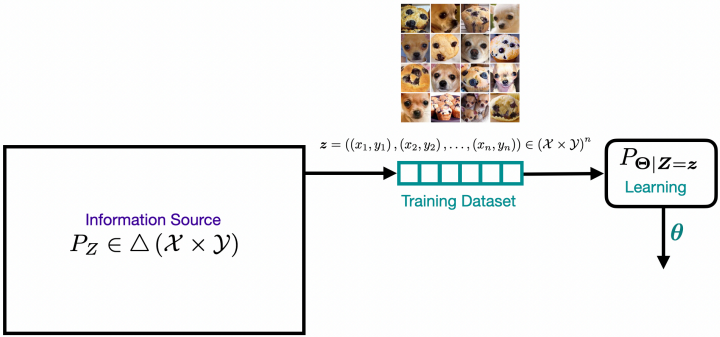

Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
Bork!
Bork!
Bork!
Bork!
Bork!
Bork!
Bork!
Bork!
Bork!

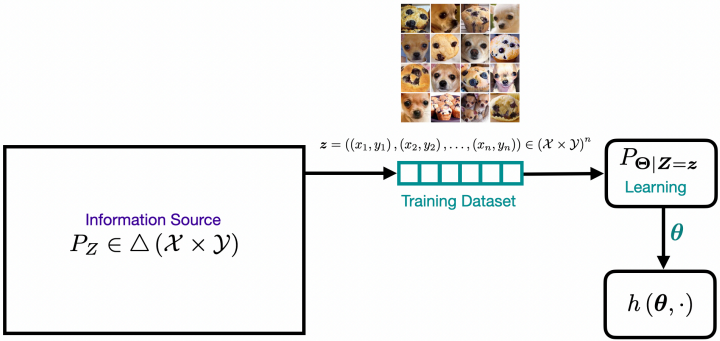

Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
$z = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Training Dataset
$P_{\Theta|Z=z}$
Learning

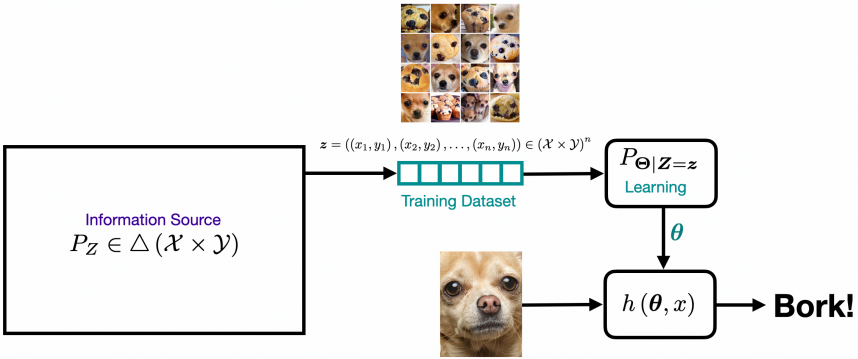


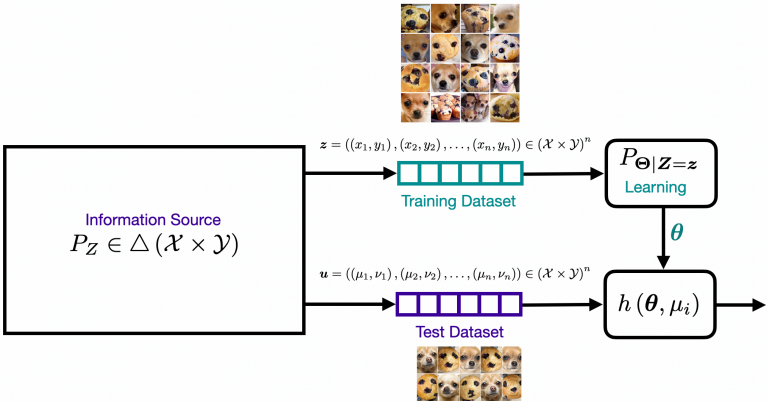
Algorithm

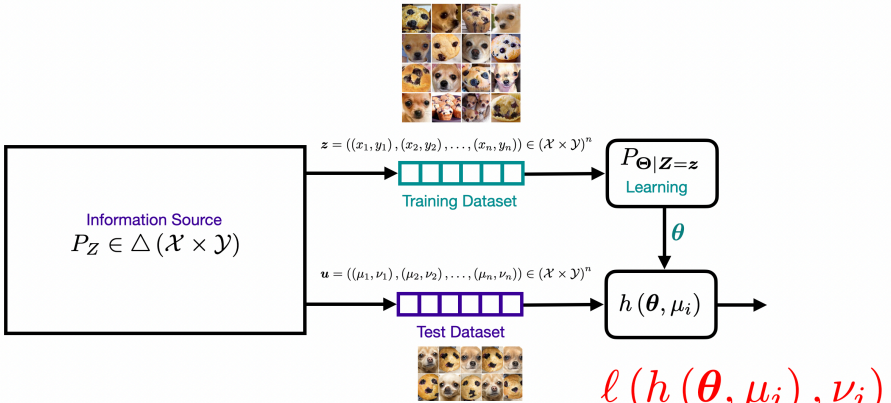
A conditional probability measure $P_{\Theta|Z} \in \triangle(\mathcal{M}|(\mathcal{X} \times \mathcal{Y})^n)$ represents a supervised machine learning algorithm.

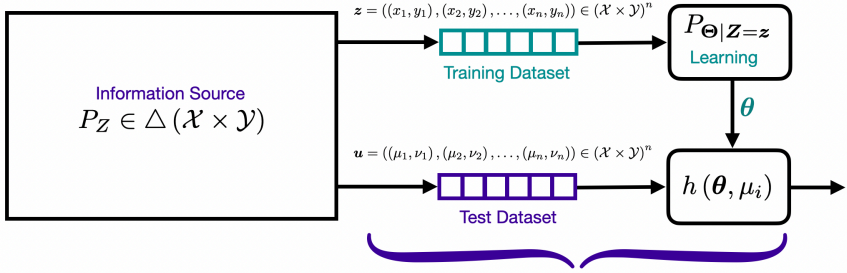

Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
$z = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Training Dataset
$P_{\Theta|Z=z}$
Learning
θ

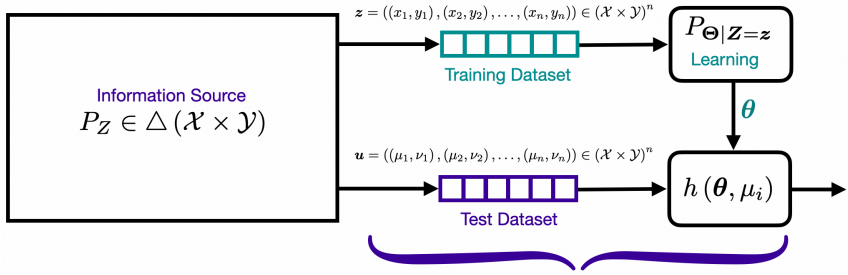

Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
$z = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Training Dataset
$P_{\Theta|Z=z}$
Learning
θ
$h(\theta, \cdot)$


Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
$z = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Training Dataset
$P_{\Theta|Z=z}$
Learning
θ
$h(\theta, x)$
Bork!


Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
$\boldsymbol{z} = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Training Dataset
$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$
Learning
$\boldsymbol{\theta}$
$\boldsymbol{u} = ((\mu_1, \nu_1), (\mu_2, \nu_2), \ldots, (\mu_n, \nu_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Test Dataset
$h(\boldsymbol{\theta}, \mu_i)$


Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
$z = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Training Dataset
$P_{\Theta|Z=z}$
Learning
θ
$u = ((\mu_1, \nu_1), (\mu_2, \nu_2), \ldots, (\mu_n, \nu_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Test Dataset
$h(\theta, \mu_i)$
$\ell(h(\theta, \mu_i), \nu_i)$


Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
$\boldsymbol{z} = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Training Dataset
$P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$
Learning
$\boldsymbol{\theta}$
$\boldsymbol{u} = ((\mu_1, \nu_1), (\mu_2, \nu_2), \ldots, (\mu_n, \nu_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Test Dataset
$h(\boldsymbol{\theta}, \mu_i)$
$\mathsf{L}(\boldsymbol{u}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^{n} \ell(h(\boldsymbol{\theta}, \mu_t), \nu_t)$

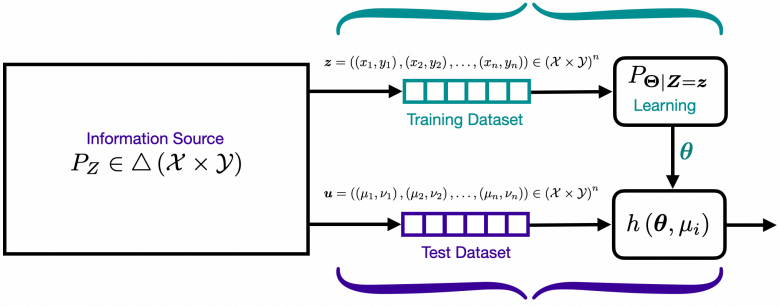


Problem Formulation: Empirical Risk Minimization (ERM)

Given the dataset $\boldsymbol{z}$, the ERM problem is

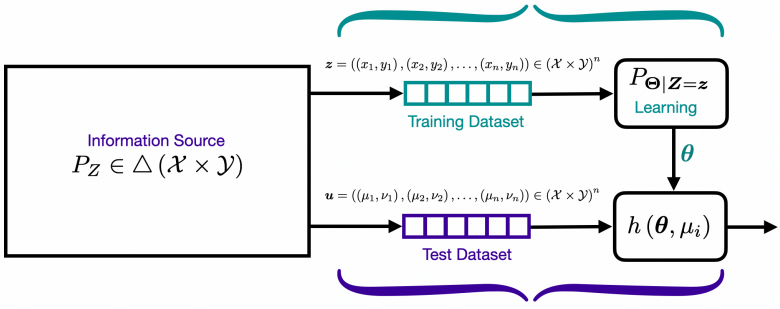
$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}).$$

$$\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\int \mathsf{L}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$



$$\mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\int \mathsf{L}\left(\boldsymbol{u},\boldsymbol{\theta}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$

$$\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\int \mathsf{L}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$



$$\mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\int \mathsf{L}\left(\boldsymbol{u},\boldsymbol{\theta}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$

Training (Expected) Risk and Test (Expected) Risk

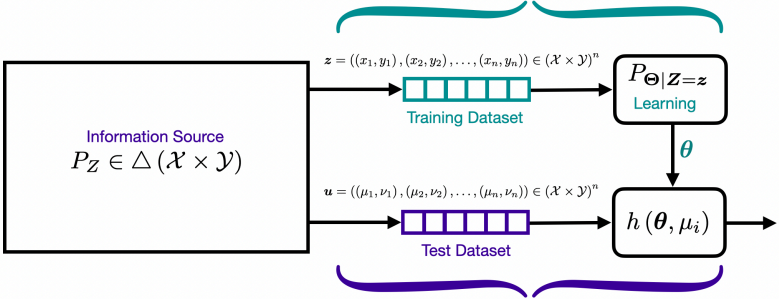
$$\underbrace{\mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}_{\text{Test Expected Risk}}-\underbrace{\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}_{\text{Training Expected Risk}}$$

Assumption:

Training datasets and **test datasets** are independent and identically distributed:

- $\boldsymbol{z}$ is drawn from $P_{\boldsymbol{Z}} \in \triangle\left((\mathcal{X} \times \mathcal{Y})^n\right)$; and
- $\boldsymbol{u}$ is drawn from $P_{\boldsymbol{Z}}$.

$$\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\int \mathsf{L}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$



$$\mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\int \mathsf{L}\left(\boldsymbol{u},\boldsymbol{\theta}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$

Generalization Error

The generalization error of the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is

$$\overline{\overline{G}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}}\right)\triangleq\int\int\left(\mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right)\mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{u}\right)\mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{z}\right).$$

ERM with Relative Entropy Regularization (ERM-RER)

Problem Formulation: ERM with Relative Entropy Regularization (ERM-RER)

The ERM-RER problem, with parameters $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $\lambda \in (0, +\infty)$, consists of the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\mathbf{z}}(P) + \lambda D(P\|Q).$$

Motivation for this regularization?

- Some priors are not probability measures:
  - Uniform distribution over infinite (countable) sets: **Counting Measure**
  - Uniform distribution over $\mathbb{R}^d$: **Lebesgue Measure**
- Some priors (probability distributions) can be calculated up to a normalization factor.
- Reference measures **constrain the set of models** $\mathcal{M}$.

S.M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical Risk Minimization with Relative Entropy Regularizations," *IEEE Trans. Inf. Theory*, vol. 70, no. 7, pp. 5122-5161, Jul. 2024.

ERM with Relative Entropy Regularization (ERM-RER)

Problem Formulation: ERM with Relative Entropy Regularization (ERM-RER)

The ERM-RER problem, with parameters $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $\lambda \in (0, +\infty)$, consists of the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(P\|Q).$$

Notation:

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp(t\,\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right) \text{ and } \mathcal{K}_{Q,\boldsymbol{z}} \triangleq \left\{s \in (0, +\infty) :\ K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) < +\infty\right\}.$$

Theorem

If $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the solution to **Problem** 1 *is unique, denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, and satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,*

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})\right).$$

S.M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical Risk Minimization with Relative Entropy Regularizations," *IEEE Trans. Inf. Theory*, vol. 70, no. 7, pp. 5122-5161, Jul. 2024.

Table of Contents

Regularization in Empirical Risk Minimization

Review of the Problem Formulation

Empirical Risk Minimization with Relative Entropy Regularization

Asymmetry of Relative Entropy on the Regularization

Support Constraint of Relative Entropy Regularization

Properties of Type-II Relative Entropy Regularization

Regularization via f-divergences

Solution and Common Regularizers

Equivalence of the f-Regularization via Transformation of the Empirical Risk

Conclusions

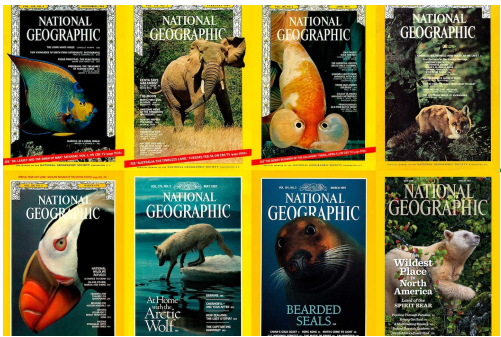
Relative Entropy Asymmetry

Definition (Generalized Relative Entropy)

Given two σ-finite measures P and Q on the same measurable space, such that $P \ll Q$

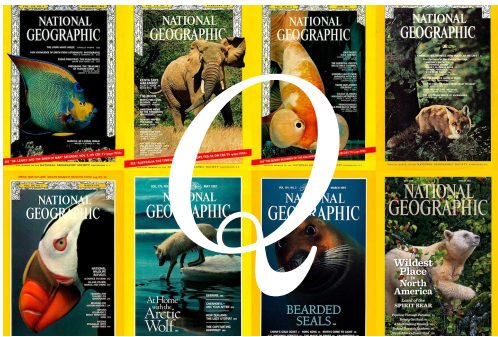
$$\mathsf{D}(P\|Q) \triangleq \int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}).$$

- Relative entropy is asymmetric: $\mathsf{D}(P\|Q) \neq \mathsf{D}(Q\|P)$
- For most cases of interest $P \ll Q \not\Longrightarrow Q \ll P$
- Solution probability measure is **constrained** to $\mathrm{supp}(P) \subseteq \mathrm{supp}(Q)$



Prior Knowledge

Prior Knowledge



Prior Knowledge

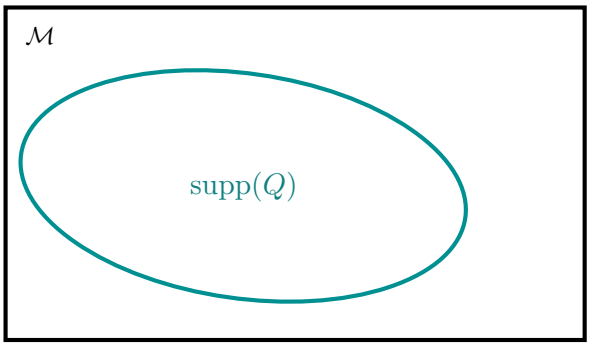
Set of All Models

$\mathcal{M}$

Prior Knowledge

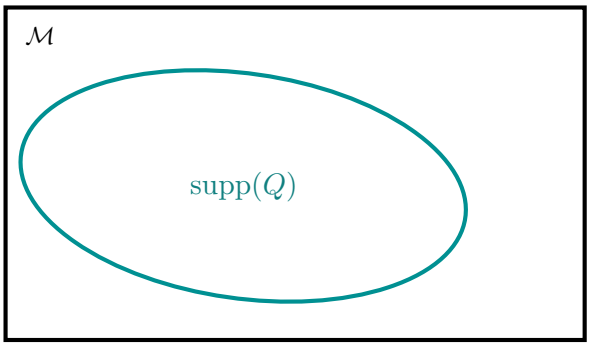
Prior Knowledge

Set of All Models

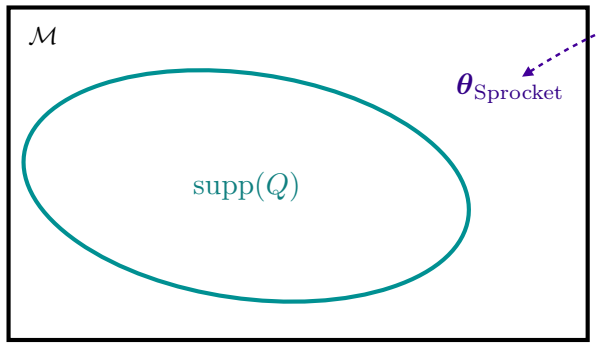


Prior Knowledge

Set of All Models

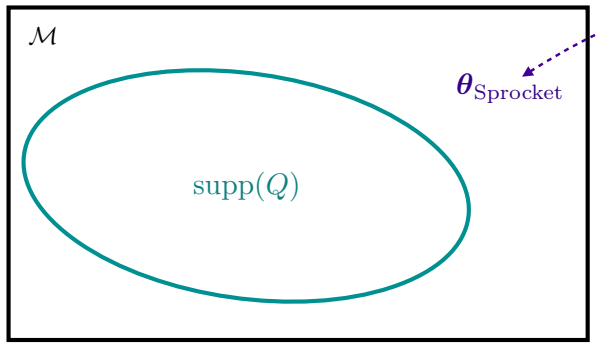


Set of All Models



Prior Knowledge

Set of All Models



Prior Knowledge

Set of All Models

$\mathcal{M}$

$\boldsymbol{\theta}_{\text{Sprocket}}$

$\text{supp}(Q)$

Prior Knowledge

$\boldsymbol{\theta}_{\text{Sprocket}} \notin \text{supp}(P)$

Type-II ERM-RER Problem

Problem Formulation: Type-II ERM-RER

The ERM-RER Type-II problem, with parameters $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $\lambda \in (0, +\infty)$, consists of the optimization over the domain $\bigtriangledown_Q(\mathcal{M}, \mathscr{F}) \triangleq \{P \in \triangle(\mathcal{M}, \mathscr{F}) : Q \ll P\}$ given by

$$\min_{P \in \bigtriangledown_Q(\mathcal{M}, \mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P).$$

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Analysis of the Relative Entropy Asymmetry in Regularized Empirical Risk Minimization," in *Proc. IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun. 2023.

Type-II ERM-RER Problem

Problem Formulation: Type-II ERM-RER

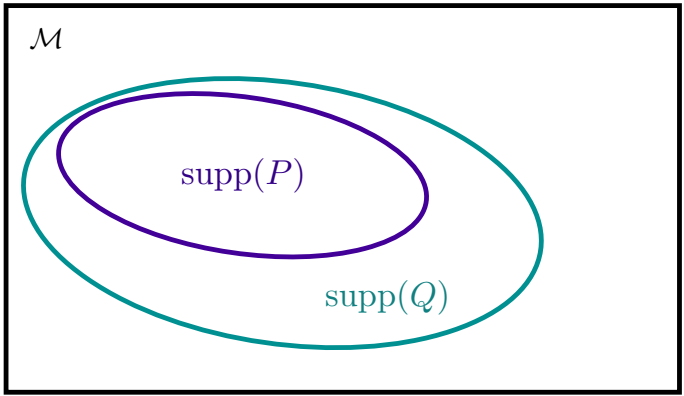
The ERM-RER Type-II problem, with parameters $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $\lambda \in (0, +\infty)$, consists of the optimization over the domain $\bigtriangledown_Q(\mathcal{M}, \mathscr{F}) \triangleq \{P \in \triangle(\mathcal{M}, \mathscr{F}) : Q \ll P\}$ given by

$$\min_{P \in \bigtriangledown_Q(\mathcal{M}, \mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P).$$

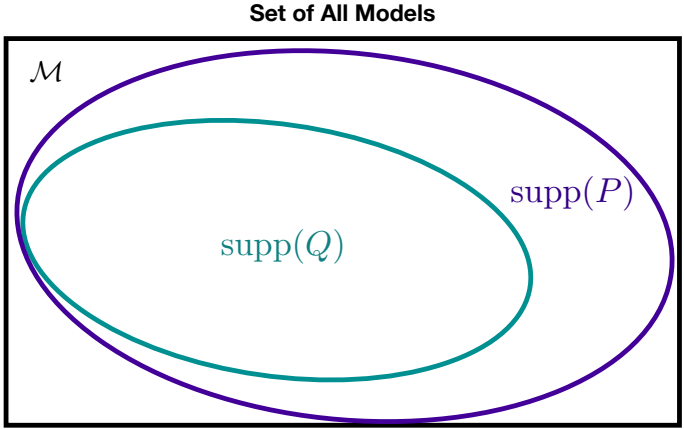
- **Asymmetry of the regularization**:
  - **Type-I ERM-RER** limits model selection to the $\mathrm{supp}(Q)$.
  - **Type-II ERM-RER** allows selection of models outside of $\mathrm{supp}(Q)$.
- Type-II regularizaiton allows exploring models outside the support of the reference

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Analysis of the Relative Entropy Asymmetry in Regularized Empirical Risk Minimization," in *Proc. IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun. 2023.

Set of All Models



Type-I Regularization: $D(P\|Q)$



Type-II Regularization: $D(Q\|P)$

Type-II ERM-RER Problem

Problem Formulation: Type-II ERM-RER with parameters Q and λ

$$\min_{P \in \bigtriangledown_Q(\mathcal{M},\mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P),$$

with $\bigtriangledown_Q(\mathcal{M},\mathscr{F}) \triangleq \{P \in \triangle(\mathcal{M},\mathscr{F}) : Q \ll P\}$

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Analysis of the Relative Entropy Asymmetry in Regularized Empirical Risk Minimization," in *Proc. IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun. 2023.

Type-II ERM-RER Problem

Problem Formulation: Type-II ERM-RER with parameters Q and λ

$$\min_{P \in \bigtriangledown_Q(\mathcal{M},\mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P),$$

with $\bigtriangledown_Q(\mathcal{M},\mathscr{F}) \triangleq \{P \in \triangle(\mathcal{M},\mathscr{F}) : Q \ll P\}$

Theorem

If there exists a real β *such that* $\beta \in \{t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < t + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})\}$ *and*

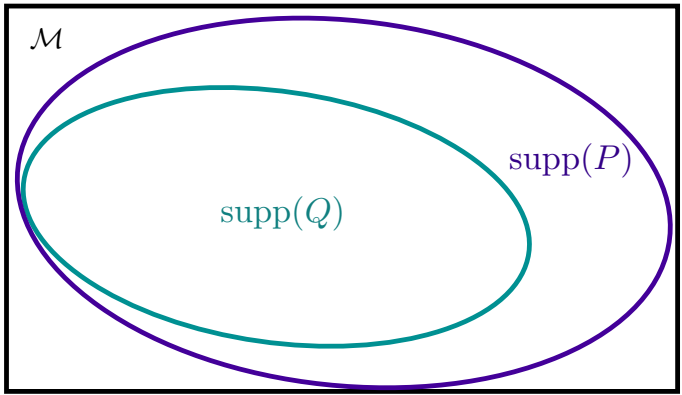
$$\int \frac{\lambda}{\beta + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})} \mathrm{d}Q(\boldsymbol{\theta}) = 1,$$

then, the unique solution to the **Type-II ERM-RER problem**, $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp}(Q)$,

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})}.$$

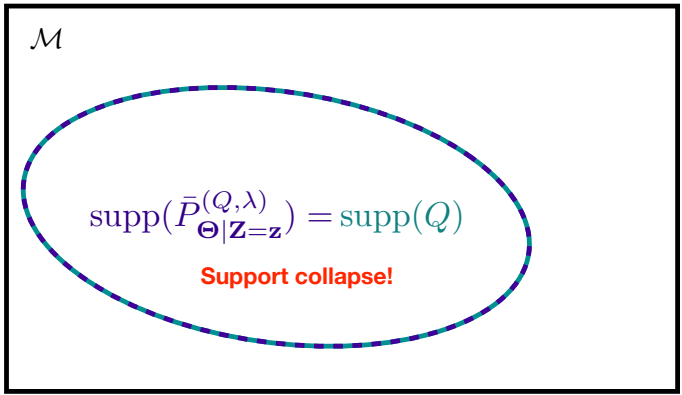
F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Analysis of the Relative Entropy Asymmetry in Regularized Empirical Risk Minimization," in *Proc. IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun. 2023.

Set of All Models



Type-II Regularization: $D(Q\|P)$

Set of All Models



Type-II Regularization: $D(Q\|P)$

Brief Sketch of the Proof:

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Analysis of the Relative Entropy Asymmetry in Regularized Empirical Risk Minimization," in *Proc. IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun. 2023.

Type-II ERM-RER Problem

Brief Sketch of the Proof:

- Solve **ancillary problem**

$$\min_{P \in \bigcirc_Q(\mathcal{M},\mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P), \quad \text{with} \quad \bigcirc_Q(\mathcal{M},\mathscr{F}) \triangleq \bigtriangledown_Q(\mathcal{M},\mathscr{F}) \cap \bigtriangleup_Q(\mathcal{M},\mathscr{F})$$

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Analysis of the Relative Entropy Asymmetry in Regularized Empirical Risk Minimization," in *Proc. IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun. 2023.

Brief Sketch of the Proof:

- Solve **ancillary problem**

$$\min_{P \in \bigcirc_Q(\mathcal{M},\mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P), \quad \text{with} \quad \bigcirc_Q(\mathcal{M},\mathscr{F}) \triangleq \bigtriangledown_Q(\mathcal{M},\mathscr{F}) \cap \bigtriangleup_Q(\mathcal{M},\mathscr{F})$$

- Show that **cost increases** outside $\bigcirc_Q(\mathcal{M},\mathscr{F})$:

$$\min_{V \in \bigtriangledown_Q(\mathcal{M},\mathscr{F}) \setminus \bigcirc_Q(\mathcal{M},\mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(V) + \lambda \mathsf{D}(Q\|V) > \min_{P \in \bigcirc_Q(\mathcal{M},\mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P)\,.$$

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Analysis of the Relative Entropy Asymmetry in Regularized Empirical Risk Minimization," in *Proc. IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun. 2023.

Brief Sketch of the Proof:

- Solve **ancillary problem**

$$\min_{P \in \bigcirc_Q(\mathcal{M},\mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P), \quad \text{with} \quad \bigcirc_Q(\mathcal{M},\mathscr{F}) \triangleq \bigtriangledown_Q(\mathcal{M},\mathscr{F}) \cap \bigtriangleup_Q(\mathcal{M},\mathscr{F})$$

- Show that **cost increases** outside $\bigcirc_Q(\mathcal{M},\mathscr{F})$:

$$\min_{V \in \bigtriangledown_Q(\mathcal{M},\mathscr{F}) \setminus \bigcirc_Q(\mathcal{M},\mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(V) + \lambda \mathsf{D}(Q\|V) > \min_{P \in \bigcirc_Q(\mathcal{M},\mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P)\,.$$

Observations:

- Type-II regularization **does not overcome induction bias** introduced by the reference measure.
- **Spoiler:** f**-divergence** regularization **does not overcome inductive bias** either.

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Analysis of the Relative Entropy Asymmetry in Regularized Empirical Risk Minimization," in *Proc. IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun. 2023.

Type-II ERM-RER Properties

Normalization Function

- The choice of λ is constrained to solutions that yield a **probability distribution**
- Let the set $\mathcal{A}_{Q,\boldsymbol{z}} \subseteq (0, \infty)$ and $\mathcal{C}_{Q,\boldsymbol{z}} \subset \mathbb{R}$ be such that if $\lambda \in \mathcal{A}_{Q,\boldsymbol{z}}$, then there exists a $\beta \in \mathcal{C}_{Q,\boldsymbol{z}}$ that satisfies $\beta \in \{t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < t + \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})\}$ and

$$\int \frac{\lambda}{\beta + \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})} \mathrm{d}Q(\boldsymbol{\theta}) = 1.$$

- The choice of λ is constrained to solutions that yield a **probability distribution**
- Let the set $\mathcal{A}_{Q,\boldsymbol{z}} \subseteq (0,\infty)$ and $\mathcal{C}_{Q,\boldsymbol{z}} \subset \mathbb{R}$ be such that if $\lambda \in \mathcal{A}_{Q,\boldsymbol{z}}$, then there exists a $\beta \in \mathcal{C}_{Q,\boldsymbol{z}}$ that satisfies $\beta \in \{t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < t + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})\}$ and

$$\int \frac{\lambda}{\beta + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})} \mathrm{d}Q(\boldsymbol{\theta}) = 1.$$

Definition (Normalization Function)

The normalization function of the Type-II ERM-RER problem is the bijection between represented by the function $\bar{K}_{Q,\boldsymbol{z}} : \mathcal{A}_{Q,\boldsymbol{z}} \to \mathcal{C}_{Q,\boldsymbol{z}}$, which satisfies $\bar{K}_{Q,\boldsymbol{z}}(\lambda) = \beta$.

Note that the Radon-Nikodym derivative of the solution is

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})}.$$

Type-II ERM-RER Properties

Optimal models without regularization

- Given a real $\delta \in [0, \infty)$, consider the set

$$\mathcal{L}_{\boldsymbol{z}}(\delta) \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \leq \delta\}.$$

- Best achievable performance **without regularization**:

$$\delta^{\star}_{Q,\boldsymbol{z}} \triangleq \inf\{\delta \in [0, \infty) : Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0\}.$$

- Solution models for the **Empirical Risk Minimization** (within $\operatorname{supp} Q$) problem:

$$\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}\}.$$

Type-II ERM-RER Properties

The Radon-Nikodym Derivative of the Solution is Positive and Finite

Type-II ERM-RER Properties

The Radon-Nikodym Derivative of the Solution is Positive and Finite

The Radon-Nikodym derivative is always finite and strictly positive.

Lemma

For all $\boldsymbol{\theta} \in \operatorname{supp} Q$ *it holds that*

$$0 < \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \leq \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} < \infty.$$

The equality holds if and only if $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$.

The Radon-Nikodym derivative is always finite and strictly positive.

Lemma

For all $\boldsymbol{\theta} \in \operatorname{supp} Q$ *it holds that*

$$0 < \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \leq \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} < \infty.$$

The equality holds if and only if $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$.

Empirical risk dominates inductive bias for any regularization regime.

Lemma

For all $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in (\operatorname{supp} Q)^2$, *such that* $\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}_1) \leq \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}_2)$, *it holds that*

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2) \leq \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1),$$

with equality if and only if $\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}_1) = \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}_2)$.

Type-II ERM-RER Properties

Asymptotes of the Radon-Nikodym Derivative

Continuity of inductive bias introduced by **large regularization factors.**

Lemma

$$\lim_{\lambda \to \infty} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 1.$$

Continuity of inductive bias introduced by **large regularization factors.**

Lemma

$$\lim_{\lambda \to \infty} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 1.$$

Continuity of inductive bias introduced by **small regularization factors.**

Lemma

If $Q(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}) > 0$ *then for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*, it holds that*

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{Q(\mathcal{L}^{\star}_{Q,\boldsymbol{z}})} \mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}\}}.$$

Type-II ERM-RER Properties

Expected Empirical Risk

Type-II ERM-RER Properties

Expected Empirical Risk

Link between expected empirical risk and normalization function:

Lemma

$$\mathsf{R}_{\boldsymbol{z}}(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda).$$

Type-II ERM-RER Properties

Expected Empirical Risk

Link between expected empirical risk and normalization function:

Lemma

$$\mathsf{R}_{\boldsymbol{z}}(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda).$$

Lower bound on the sensitivity of $\mathsf{R}_{\boldsymbol{z}}$:

Lemma

$$\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}) \geq \lambda(\exp(\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)) - 1).$$

Link between expected empirical risk and normalization function:

Lemma

$$\mathsf{R}_{\boldsymbol{z}}(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda).$$

Lower bound on the sensitivity of $\mathsf{R}_{\boldsymbol{z}}$:

Lemma

$$\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}) \geq \lambda(\exp(\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)) - 1).$$

Bounds on the **expected empirical risk**:

Lemma

$$\delta^{\star}_{Q,\boldsymbol{z}} \leq \mathsf{R}_{\boldsymbol{z}}(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}) < \lambda + \delta^{\star}_{Q,\boldsymbol{z}}.$$

Equality holds if and only if the empirical risk function is nonseparable.

Equilavence of **Type-I** and **Type-II** Regularization

Theorem

Type-II $\Rightarrow$ Type-I Equivalence:

$$\min_{P\in\nabla_Q(\mathcal{M})} \int \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})\mathrm{d}P(\boldsymbol{\theta}) \ + \lambda \mathsf{D}(Q\|P) \ = \min_{P\in\triangle_Q(\mathcal{M})} \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\mathrm{d}P(\boldsymbol{\theta}) \ + \mathsf{D}(P\|Q)\,,$$

where the function $\mathsf{V}_{Q,\boldsymbol{z},\lambda}\mathcal{M}\to\mathbb{R}$*, referred to as the log-empirical risk, is defined as*

$$\mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) = \log(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})).$$

Type-I $\Rightarrow$ Type-II Equivalence:

$$\min_{P\in\triangle_Q(\mathcal{M})} \int \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})\mathrm{d}P(\boldsymbol{\theta}) \ + \lambda \mathsf{D}(P\|Q) \ = \min_{P\in\nabla_Q(\mathcal{M})} \int \mathsf{W}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\mathrm{d}P(\boldsymbol{\theta}) \ + \mathsf{D}(Q\|P)\,,$$

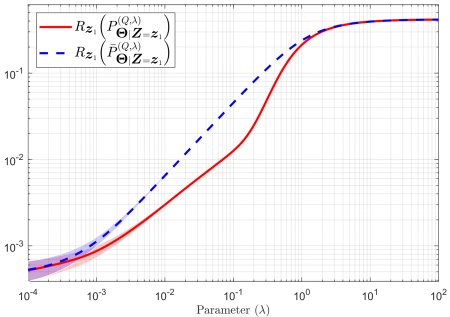
where the function $\mathsf{W}_{Q,\boldsymbol{z},\lambda}:\mathcal{M}\to\mathbb{R}$ *is defined as*

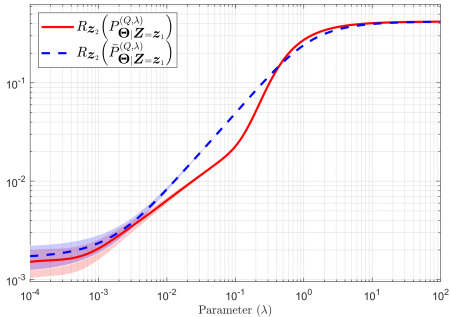
$$\mathsf{W}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) = \frac{\lambda}{\exp(-\frac{\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})}{\lambda} - K_{Q,\boldsymbol{z}}(-\frac{1}{\lambda}))} - \bar{K}_{Q,\boldsymbol{z}}(\lambda).$$

Numerical Comparison of **Type-I** and **Type-II** Regularization

Evaluation of the Generalization Capabilities

We train a **binary classifier** to distinguish `'six'` and `'seven'` in the MNIST dataset with the ERM-RER Type-I and Type-II

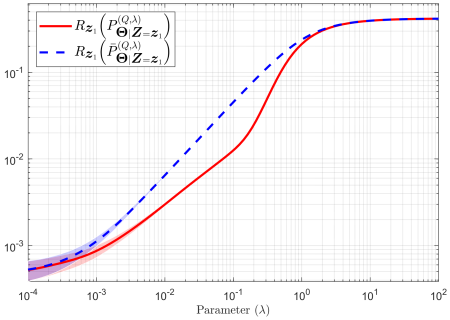




Numerical Comparison of **Type-I** and **Type-II** Regularization

Evaluation of the Generalization Capabilities

We train a **binary classifier** to distinguish ‘six’ and ‘seven’ in the MNIST dataset with the ERM-RER Type-I and Type-II



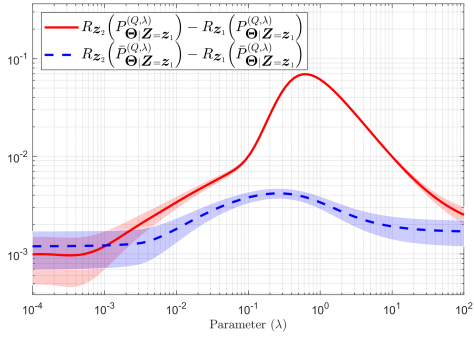


Table of Contents

Regularization in Empirical Risk Minimization

Review of the Problem Formulation

Empirical Risk Minimization with Relative Entropy Regularization

Asymmetry of Relative Entropy on the Regularization

Support Constraint of Relative Entropy Regularization

Properties of Type-II Relative Entropy Regularization

Regularization via f-divergences

Solution and Common Regularizers

Equivalence of the f-Regularization via Transformation of the Empirical Risk

Conclusions

Definition (f-divergence [Csiszár, 1967])

Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Let P and Q be two probability measures on the measurable space $(\mathcal{M}, \mathscr{F})$. If the probability measure P is absolutely continuous with respect to the probability measure Q then the f-divergence is defined as

$$\mathsf{D}_f(P\|Q) \triangleq \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}),$$

where $f(0) = \lim_{x \to 0^+} f(x)$.

Information-type measures of dissimilarity between two probability distributions [Csiszár, 1967].

Motivation and significance:

- Operational insight in:
  - Channel coding
  - Compression, estimation
  - High-dimensional statistics
  - Hypothesis testing
- Amenable to variational representations
- Link to *Fisher information*

Common f-divergences:

- Relative Entropy: $f(x) = x \log x$
- Total Variation: $f(x) = \frac{1}{2}|x-1|$
- χ^2-divergence: $f(x) = (x-1)^2$
- Squared Hellinger distance: $f(x) = (1-\sqrt{x})^2$
- Jensen-Shannon divergence: $f(x) = x \log\left(\frac{2x}{x+1}\right) + \log\left(\frac{2}{x+1}\right)$

Basic Properties

- $\mathsf{D}_f(P\|P) = 0$.
- $\mathsf{D}_f(P\|Q) \geq 0$. If f is strictly convex then $\mathsf{D}_f(P\|Q) = 0 \iff P = Q$.
- $\mathsf{D}_f(P_{X,Y}\|Q_{X,Y}) \geq \mathsf{D}_f(P_X\|Q_X)$.
- $(P, Q) \mapsto \mathsf{D}_f(P\|Q)$ is jointly convex.
  - $P \mapsto \mathsf{D}_f(P\|Q)$ is convex
  - $Q \mapsto \mathsf{D}_f(P\|Q)$ is convex

Problem Formulation: ERM with f-divergence Regularization (ERM-fDR)

Given the dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, the ERM-$f$DR problem, with parameters Q , λ, and f, consists of the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{F})} \quad \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}_f(P \| Q)\,,$$

with optimization domain

$$\triangle_Q\left(\mathcal{M}, \mathscr{F}\right) \triangleq \{P \in \triangle(\mathcal{M}, \mathscr{F}) : P \ll Q\}.$$

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Equivalence of the Empirical Risk Minimization to Regularization on the Family of f-Divergences,," in *Proc. IEEE International Symposium on Information Theory*, Athens, Greece, Jul. 2024.

ERM wih f-divergence Regularization

Assumptions

- The function f is strictly **convex** and **differentiable**
- There exists a β such that

$$\beta \in \left\{ t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < \dot{f}^{-1}\left(-\frac{t + \mathsf{L}\left(\boldsymbol{z}, \boldsymbol{\theta}\right)}{\lambda} \right) \right\}$$

  and

$$\int \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})}{\lambda} \right) \mathrm{d}Q(\boldsymbol{\theta}) = 1$$

- The function $\mathsf{L}_{\boldsymbol{z}}$ is **separable** with respect to the probability measure Q

- The function f is strictly **convex** and **differentiable**
- There exists a β such that

$$\beta \in \left\{ t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < \dot{f}^{-1}\left(-\frac{t + \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})}{\lambda} \right) \right\}$$

and

$$\int \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})}{\lambda} \right) \mathrm{d}Q(\boldsymbol{\theta}) = 1$$

- The function $\mathsf{L}_{\boldsymbol{z}}$ is **separable** with respect to the probability measure Q

Definition (Separable Empirical Risk Function)

The empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ is said to be separable with respect to a σ-finite measure $P \in \triangle(\mathcal{M})$, if there exist a positive real $c > 0$ and two subsets $\mathcal{A}$ and $\mathcal{C}$ of $\mathcal{M}$ that are nonneglible with respect to P, such for all $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathcal{A} \times \mathcal{C}$, it holds that

$$\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}_1) < c < \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}_2) < \infty.$$

Theorem

Under assumptions stated in the previous slide, the solution to the ERM-fDR problem is unique, and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, is given by

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})}{\lambda}\right).$$

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Equivalence of the Empirical Risk Minimization to Regularization on the Family of f-Divergences,," in *Proc. IEEE International Symposium on Information Theory*, Athens, Greece, Jul. 2024.

Theorem

Under assumptions stated in the previous slide, the solution to the ERM-fDR problem is unique, and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, is given by

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})}{\lambda}\right).$$

Remarks:

- Probability measures Q and $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ are **mutually absolutely continuous**.
- **No support exploration:** f-divergence regularization forces the solution to coincide with the support of the reference measure Q, independently of the training data.

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Equivalence of the Empirical Risk Minimization to Regularization on the Family of f-Divergences,," in *Proc. IEEE International Symposium on Information Theory*, Athens, Greece, Jul. 2024.

ERM wih f-divergence Regularization

Common Cases: Kullback-Leibler Divergence (Type-I)

Setting

$$f(x) = x \log x,$$
$$\dot{f}(x) = \log x + 1,$$

results in

$$\mathsf{D}_f(P\|Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P(\boldsymbol{\theta}).$$

The ERM-fDR solution yields

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-\frac{\beta + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta}) + \lambda}{\lambda}\right).$$

ERM wih f-divergence Regularization

Common Cases: Kullback-Leibler Divergence (Type-II)

Setting

$$f(x) = -\log x,$$
$$\dot{f}(x) = -\frac{1}{x},$$

results in

$$\mathsf{D}_f(P\|Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}) = -\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}) = \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}).$$

The ERM-fDR solution yields

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\beta + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})}.$$

ERM wih f-divergence Regularization

Common Cases: Jensen-Shannon Divergence

Definition (Jensen-Shannon Divergence)

Let P and Q be two probability measures on the measurable space $(\mathcal{M}, \mathscr{F})$. If the probability measure P is absolutely continuous with respect to the probability measure Q then the Jensen-Shannon divergence is

$$\mathrm{JS}\left(P, Q\right) = \mathsf{D}\left(P \| \frac{1}{2}(P+Q)\right) + \mathsf{D}\left(Q \| \frac{1}{2}(P+Q)\right).$$

- **Remark:** $\sqrt{\mathrm{JS}\left(P, Q\right)}$ is a metric in the space of probability measure.
- The link to f-divergence characterization is

$$f(x) = x \log\left(\frac{2x}{x+1}\right) + \log\left(\frac{2}{x+1}\right),$$
$$\dot{f}(x) = \log\left(\frac{2x}{x+1}\right).$$

- The ERM-fDR solution yields

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{2\exp(\frac{\beta+\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})}{\lambda}) - 1}.$$

ERM wih f-divergence Regularization

Common Cases: χ^2-divergence

Definition (χ^2-divergence)

Let P and Q be two probability measures on the measurable space $(\mathcal{M}, \mathscr{F})$. If the probability measure P is absolutely continuous with respect to the probability measure Q then the χ^2-divergence is

$$\chi^2\left(P\|Q\right) = \frac{1}{2}\int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) - 1\right)^2 \mathrm{d}Q(\boldsymbol{\theta}).$$

- The link to f-divergence characterization is

$$f(x) = (x-1)^2,$$
$$\dot{f}(x) = 2(x-1).$$

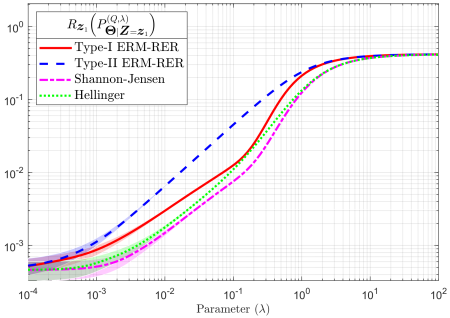
- The ERM-fDR solution yields

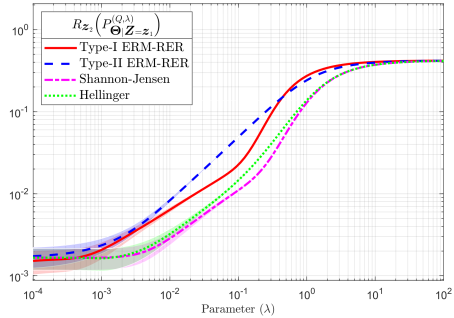
$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = -\frac{\beta + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})}{\lambda}.$$

Numerical Comparison of Several Regularizations

Evaluation of the Generalization Capabilities

We train a **binary classifier** to distinguish ‘six’ and ‘seven’ in the MNIST dataset with the ERM-RER **several regularizers**.

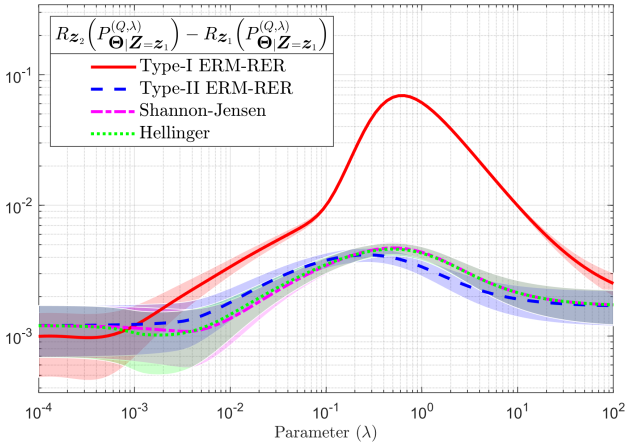




Numerical Comparison of Several Regularizations

Evaluation of the Generalization Capabilities

We train a **binary classifier** to distinguish `'six'` and `'seven'` in the MNIST dataset with the ERM-RER **several regularizers**.



Revisiting the Regularization Equivalence

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Equivalence of the Empirical Risk Minimization to Regularization on the Family of f-Divergences,," in *Proc. IEEE International Symposium on Information Theory*, Athens, Greece, Jul. 2024.

Revisiting the Regularization Equivalence

- Recall that **Type-I** and **Type-II** regularizations are **equivalent via a transformation** of the expected empirical risk: **does this extend to f-divergence regularization?**

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Equivalence of the Empirical Risk Minimization to Regularization on the Family of f-Divergences,," in *Proc. IEEE International Symposium on Information Theory*, Athens, Greece, Jul. 2024.

Revisiting the Regularization Equivalence

- Recall that **Type-I** and **Type-II** regularizations are **equivalent via a transformation** of the expected empirical risk: **does this extend to f-divergence regularization?**

Theorem

Let f and g be two strictly convex and differentiable functions satisfying the conditions to generate an f-divergence and g-divergence, respectively. If the following problem possess solutions, then

$$\min_{P\in\triangle_Q(\mathcal{M})} \int \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})\mathrm{d}P(\boldsymbol{\theta}) \; + \lambda \mathsf{D}_f(P\|Q) \;\; = \min_{P\in\triangle_Q(\mathcal{M})} \int v(\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta}))\mathrm{d}P(\boldsymbol{\theta}) \; + \lambda \mathsf{D}_g(P\|Q),$$

where the function $v:[0,\infty)\to\mathbb{R}$ is such that

$$v(t) = \lambda \dot{g}\left(\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+t}{\lambda}\right)\right) - N'_{Q,\boldsymbol{z}}(\lambda),$$

with $N_{Q,\boldsymbol{z}}$ and $N'_{Q,\boldsymbol{z}}$ being the respective normalization functions.

F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Equivalence of the Empirical Risk Minimization to Regularization on the Family of f-Divergences,," in *Proc. IEEE International Symposium on Information Theory*, Athens, Greece, Jul. 2024.

Table of Contents

Regularization in Empirical Risk Minimization

Review of the Problem Formulation

Empirical Risk Minimization with Relative Entropy Regularization

Asymmetry of Relative Entropy on the Regularization

Support Constraint of Relative Entropy Regularization

Properties of Type-II Relative Entropy Regularization

Regularization via f-divergences

Solution and Common Regularizers

Equivalence of the f-Regularization via Transformation of the Empirical Risk

Conclusions

Conclusions for Part III

- **All** f-Divergence regularizations to the ERM problem exhibit solutions that are **mutually absolutely continuous** with the reference measure.
  - What implications on the set of models that would exhibit **positive probability**?
  - How to choose Q ?
- Several solutions to the ERM-fDR problem simultaneously exhibit **smaller training and test errors** than those induced by the **Gibbs Algorithm**.
- Equivalence results for f-divergence regularization unveil **link** between the choice of f-divergence and **loss** function.

Conclusions for Part III

- **All** f-Divergence regularizations to the ERM problem exhibit solutions that are **mutually absolutely continuous** with the reference measure.
  - What implications on the set of models that would exhibit **positive probability**?
  - How to choose Q ?
- Several solutions to the ERM-fDR problem simultaneously exhibit **smaller training and test errors** than those induced by the **Gibbs Algorithm**.
- Equivalence results for f-divergence regularization unveil **link** between the choice of f-divergence and **loss** function.
- **Adapting** the f-divergence to different learning frameworks suggests **tailored regularizer designs**
  - Loss function definition
  - Model set adaptation to practical implementations

Conclusions for Part III

- **All** f-Divergence regularizations to the ERM problem exhibit solutions that are **mutually absolutely continuous** with the reference measure.
  - What implications on the set of models that would exhibit **positive probability**?
  - How to choose Q ?
- Several solutions to the ERM-fDR problem simultaneously exhibit **smaller training and test errors** than those induced by the **Gibbs Algorithm**.
- Equivalence results for f-divergence regularization unveil **link** between the choice of f-divergence and **loss** function.
- **Adapting** the f-divergence to different learning frameworks suggests **tailored regularizer designs**
  - Loss function definition
  - Model set adaptation to practical implementations
- **Open problem:** How to choose *all* these parameters λ, Q, f, ℓ, ...

Bibliography I

Csiszár, I. (1967).

Information-type measures of difference of probability distributions and indirect observation.

Studia Scientiarum Mathematicarum Hungarica, 2(1):299–318.

Slides for Part IV

Table of Contents (Part IV)

Empirical Risk **Optimization** with Relative Entropy Regularization

The Method of Gaps

Explicit Expressions for the Generalization Error

Concluding Remarks

Table of Contents

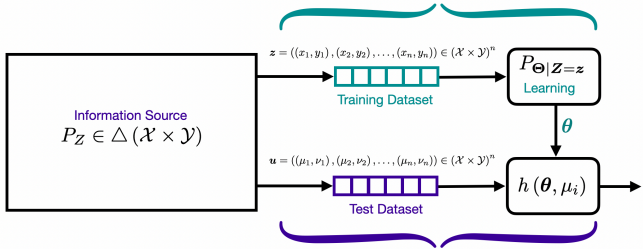
Empirical Risk **Optimization** with Relative Entropy Regularization

The Method of Gaps

Explicit Expressions for the Generalization Error

Concluding Remarks

$$\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\int \mathsf{L}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$



$$\mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\int \mathsf{L}\left(\boldsymbol{u},\boldsymbol{\theta}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$

Generalization Error (Definition 4 in [Perlaza-2024b])

The generalization error of the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is

$$\overline{\overline{G}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}}\right)\triangleq\int\int\left(\mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right)\mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{u}\right)\mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{z}\right).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Empirical Risk **Minimization** with Relative Entropy Regularization

The Gibbs Algorithm

- Given a **fixed dataset** $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$; and
- given a **reference measure** $Q \in \triangle(\mathcal{M})$ and a **real** $\lambda > 0$

Problem 1: ERM with Relative Entropy Regularization

$$\min_{P \in \triangle_Q(\mathcal{M})} \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \lambda \mathsf{D}(P \| Q),$$

with $\triangle_Q(\mathcal{M}) \triangleq \{P \in \triangle(\mathcal{M}) : P \ll Q\}$.

Empirical Risk **Minimization** with Relative Entropy Regularization

The Gibbs Algorithm

- Given a **fixed dataset** $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$; and
- given a **reference measure** $Q \in \triangle(\mathcal{M})$ and a **real** $\lambda > 0$

Problem 1: ERM with Relative Entropy Regularization

$$\min_{P \in \triangle_Q(\mathcal{M})} \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \lambda \mathsf{D}(P \| Q),$$

with $\triangle_Q(\mathcal{M}) \triangleq \{P \in \triangle(\mathcal{M}) : P \ll Q\}$.

Problem 1*a*: ERM within a Neighborhood

$$\min_{P \in \triangle_Q(\mathcal{M})} \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta})$$
$$\text{s.t.} \quad \mathsf{D}(P \| Q) \leqslant \gamma.$$

[Perlaza-2024a] Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, and Stefano Rini. "**Empirical Risk Minimization with Relative Entropy Regularization**". IEEE Transactions on Information Theory, vol. 70, no. 7, pp. 5122 – 5161, July, 2024.

Empirical Risk **Maximization** with Relative Entropy Regularization

Worst-Case Data-Generating Probability Measure

- Given a **fixed model** $\boldsymbol{\theta} \in \mathcal{M}$; and
- Given a **reference measure** $P_S \in \triangle(\mathcal{X} \times \mathcal{Y})$ and **a real** $\beta > 0$

Problem 2: Loss Maximization with Relative Entropy Regularization

$$\max_{P \in \triangle_{P_S}(\mathcal{X} \times \mathcal{Y})} \int \ell(h(\boldsymbol{\theta}, x), y) \, \mathrm{d}P(x, y) - \beta \mathrm{D}(P \| P_S),$$

with $\triangle_{P_S}(\mathcal{X} \times \mathcal{Y}) \triangleq \{P \in \triangle(\mathcal{X} \times \mathcal{Y}) : P \ll P_S\}$.

Empirical Risk **Maximization** with Relative Entropy Regularization

Worst-Case Data-Generating Probability Measure

- Given a **fixed model** $\boldsymbol{\theta} \in \mathcal{M}$; and
- Given a **reference measure** $P_S \in \triangle(\mathcal{X} \times \mathcal{Y})$ and **a real** $\beta > 0$

Problem 2: Loss Maximization with Relative Entropy Regularization

$$\max_{P \in \triangle_{P_S}(\mathcal{X} \times \mathcal{Y})} \int \ell\left(h\left(\boldsymbol{\theta}, x\right), y\right) \mathrm{d}P(x, y) - \beta \mathsf{D}(P \| P_S),$$

with $\triangle_{P_S}(\mathcal{X} \times \mathcal{Y}) \triangleq \{P \in \triangle(\mathcal{X} \times \mathcal{Y}) : P \ll P_S\}$.

Problem 2: Loss Maximization within a Neighbourhood

$$\max_{P \in \triangle_{P_S}(\mathcal{X} \times \mathcal{Y})} \int \ell\left(h\left(\boldsymbol{\theta}, x\right), y\right) \mathrm{d}P(x, y)$$
$$\text{s.t.} \qquad \mathsf{D}(P \| P_S) \leqslant \gamma.$$

[Zou-2024] Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman, and H. Vincent Poor. "**The Worst-Case Data-Generating Probability Measure in Statistical Learning**". IEEE Journal on Selected Areas in Information Theory, vol. 5, pp. 175–189, Apr., 2024.

Problem 1:

Empirical Risk **Minimization** with Relative Entropy Regularization

Empirical Risk **Minimization** with Relative Entropy Regularization

Problem 1: ERM with Relative Entropy Regularization

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \lambda \mathsf{D}(P \| Q),$$

with $\triangle_Q(\mathcal{M}) \triangleq \{P \in \triangle(\mathcal{M}) : P \ll Q\}$.

Notation:

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp(t\,\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right) \text{ and } \mathcal{K}_{Q,\boldsymbol{z}} \triangleq \left\{s \in (0, +\infty) : \; K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) < +\infty\right\}$$

Theorem (Theorem 3 in [Perlaza-2024a])

If $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the solution to **Problem 1** *is unique, denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, and satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,*

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})\right).$$

[Perlaza-2024a] Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, and Stefano Rini. "**Empirical Risk Minimization with Relative Entropy Regularization**". IEEE Transactions on Information Theory, vol. 70, no. 7, pp. 5122 – 5161, July, 2024.

Empirical Risk **Minimization** with Relative Entropy Regularization

Problem 1: ERM with Relative Entropy Regularization

$$\min_{P \in \triangle_Q(\mathcal{M})} \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \lambda \mathsf{D}(P \| Q),$$

with $\triangle_Q(\mathcal{M}) \triangleq \{P \in \triangle(\mathcal{M}) : P \ll Q\}$.

Notation:

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp(t\,\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta})\right) \text{ and } \mathcal{K}_{Q,\boldsymbol{z}} \triangleq \left\{s \in (0, +\infty) : \; K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) < +\infty\right\}$$

Theorem (Equation (28) in [Perlaza-2024a])

If $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the solution to **Problem 1** *is a unique, denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, and satisfies for all $\boldsymbol{\theta} \in \operatorname{sup} Q$,*

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{1}{\lambda}\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})\right)}{\int \exp\left(-\frac{1}{\lambda}\,\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})}.$$

[Perlaza-2024a] Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, and Stefano Rini. "**Empirical Risk Minimization with Relative Entropy Regularization**". IEEE Transactions on Information Theory, vol. 70, no. 7, pp. 5122 – 5161, July, 2024.

Empirical Risk **Minimization** with Relative Entropy Regularization

Problem 1: ERM with Relative Entropy Regularization

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \underbrace{\int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta})}_{\mathsf{R}_{\boldsymbol{z}}(P)} + \lambda \mathsf{D}(P \| Q) .$$

Solution:

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})\right).$$

Sensitivity to **deviations from the Optimal Measure**:

Lemma (Lemma 33 in [Perlaza-2024b])

$$\mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \lambda\left(\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) + \mathsf{D}\left(P \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}(P \| Q)\right)$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Empirical Risk **Minimization** with Relative Entropy Regularization

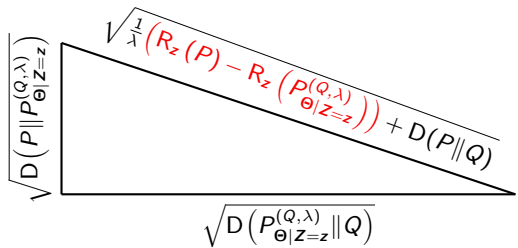


Figure: Geometric interpretation of the gap $R_z(P) - R_z\left(P^{(Q,\lambda)}_{\Theta|Z=z}\right)$.

Empirical Risk **Minimization** with Relative Entropy Regularization

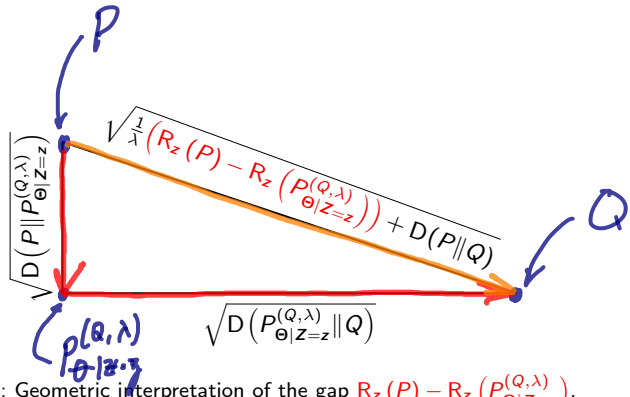


Figure: Geometric interpretation of the gap $R_z(P) - R_z\left(P^{(Q,\lambda)}_{\Theta|Z=z}\right)$.

Empirical Risk **Minimization** with Relative Entropy Regularization

Problem 1: ERM with Relative Entropy Regularization

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \underbrace{\int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta})}_{\mathsf{R}_{\boldsymbol{z}}(P)} + \lambda \mathsf{D}(P \| Q) \, .$$

Solution:

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})\right).$$

Theorem (Theorem 37 in [Perlaza-2024b])

For all $P_1 \in \triangle_Q(\mathcal{M})$ and $P_2 \in \triangle_Q(\mathcal{M})$,

$$\mathsf{R}_{\boldsymbol{z}}(P_1) - \mathsf{R}_{\boldsymbol{z}}(P_2) = \lambda\left(\mathsf{D}\left(P_1 \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}\left(P_2 \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}(P_2 \| Q) - \mathsf{D}(P_1 \| Q)\right).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Problem 2:

Loss **Maximization** with Relative Entropy Regularization

Loss **Maximization** with Relative Entropy Regularization

Problem 2: Loss Maximization with Relative Entropy Regularization

$$\max_{P \in \triangle_{P_S}(\mathcal{X}\times\mathcal{Y})} \int \ell\left(h\left(\boldsymbol{\theta}, x\right), y\right) \mathrm{d}P(x,y) - \beta \mathsf{D}(P\|P_S),$$

with $\triangle_{P_S}(\mathcal{X}\times\mathcal{Y}) \triangleq \{P \in \triangle(\mathcal{X}\times\mathcal{Y}) : P \ll P_S\}$.

Notation:

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}(t) = \log\left(\int \exp\left(t\ell(\boldsymbol{\theta}, x, y)\right)\mathrm{d}P_S(x,y)\right) \text{ and } \mathcal{J}_{P_S,\boldsymbol{\theta}} \triangleq \left\{t \in (0,+\infty) : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) < +\infty\right\}$$

Theorem (Theorem 1 in [Zou-2024])

If $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, the solution to **Problem** 2 *is unique, denoted by $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$, and satisfies for all $(x,y) \in \operatorname{supp} P_S$,*

$$\frac{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x,y) = \exp\left(\frac{1}{\beta}\ell\left(h\left(\boldsymbol{\theta}, x\right), y\right) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right).$$

[Zou-2024] Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman, and H. Vincent Poor. "**The Worst-Case Data-Generating Probability Measure in Statistical Learning**". IEEE Journal on Selected Areas in Information Theory, vol. 5, pp. 175–189, Apr., 2024.

Loss **Maximization** with Relative Entropy Regularization

Problem 2: Loss Maximization with Relative Entropy Regularization

$$\max_{P \in \triangle_{P_S}(\mathcal{X}\times\mathcal{Y})} \underbrace{\int \ell\left(h\left(\boldsymbol{\theta}, x\right), y\right) \mathrm{d}P(x,y)}_{\mathsf{R}_{\boldsymbol{\theta}}(P)} - \beta \mathsf{D}\left(P \| P_S\right),$$

Solution:

$$\frac{\mathrm{d}P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x,y) = \exp\left(\frac{1}{\beta}\ell\left(h\left(\boldsymbol{\theta}, x\right), y\right) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right).$$

Assumption: $P_{\boldsymbol{Z}}\left(\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n\right) = \prod_{t=1}^{n} P_Z\left(\mathcal{A}_t\right)$

Lemma (Theorem 6 in [Zou-2024])

$$\mathsf{R}_\theta\left(P\right) - \mathsf{R}_\theta\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) = \beta\left(\mathsf{D}(P\|P_S) - \mathsf{D}\left(P\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - \mathsf{D}\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right)\right)$$

[Zou-2024] Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman, and H. Vincent Poor. "**The Worst-Case Data-Generating Probability Measure in Statistical Learning**". IEEE Journal on Selected Areas in Information Theory, vol. 5, pp. 175–189, Apr., 2024.

Loss **Maximization** with Relative Entropy Regularization

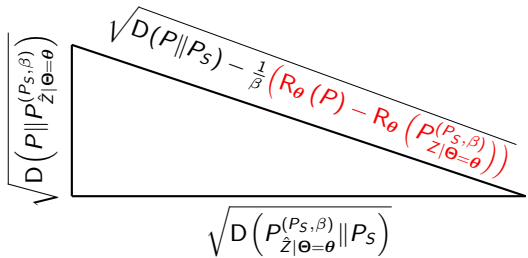


Figure: Geometric interpretation of the gap $\mathsf{R}_{\boldsymbol{\theta}}(P) - \mathsf{R}_{\boldsymbol{\theta}}\left(P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$.

Loss **Maximization** with Relative Entropy Regularization

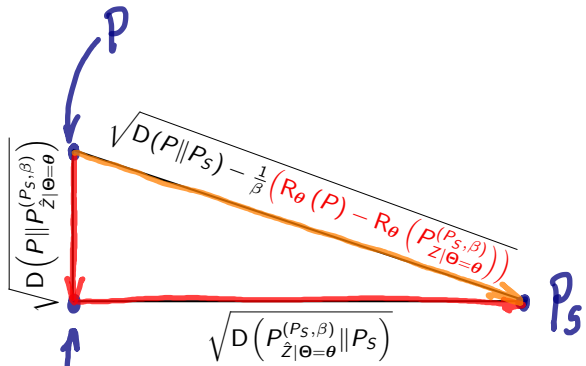


Figure: Geometric interpretation of the gap $\mathrm{R}_{\theta}(P) - \mathrm{R}_{\theta}\left(P^{(P_S,\beta)}_{Z|\Theta=\theta}\right)$.

Loss **Maximization** with Relative Entropy Regularization

Problem 2: Loss Maximization with Relative Entropy Regularization

$$\max_{P \in \triangle_{P_S}(\mathcal{X}\times\mathcal{Y})} \underbrace{\int \ell\left(h\left(\boldsymbol{\theta}, x\right), y\right) \mathrm{d}P(x,y)}_{\mathsf{R}_{\boldsymbol{\theta}}(P)} - \beta \mathsf{D}\left(P \| P_S\right),$$

Solution:

$$\frac{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x,y) = \exp\left(\frac{1}{\beta}\ell\left(h\left(\boldsymbol{\theta}, x\right), y\right) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right).$$

Assumption: $P_{\boldsymbol{Z}}\left(\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n\right) = \prod_{t=1}^{n} P_Z\left(\mathcal{A}_t\right)$

Theorem (Theorem 8 in [Zou-2024])

For all $P_1 \in \triangle_{P_S}\left(\mathcal{X}\times\mathcal{Y}\right)$ *and for all* $P_2 \in \triangle_{P_S}\left(\mathcal{X}\times\mathcal{Y}\right)$,

$$\mathsf{R}_{\boldsymbol{\theta}}\left(P_1\right) - \mathsf{R}_{\boldsymbol{\theta}}\left(P_2\right) = \beta\Big(\mathsf{D}\left(P_2 \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - \mathsf{D}\left(P_1 \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - \mathsf{D}(P_2\|P_S) + \mathsf{D}(P_1\|P_S)\Big)$$

[Zou-2024] Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman, and H. Vincent Poor. "**The Worst-Case Data-Generating Probability Measure in Statistical Learning**". IEEE Journal on Selected Areas in Information Theory, vol. 5, pp. 175–189, Apr., 2024.

So far...

Theorem (Theorem 37 in [Perlaza-2024b])

For all $P_1 \in \triangle_Q(\mathcal{M})$ *and* $P_2 \in \triangle_Q(\mathcal{M})$,

$$\mathsf{R}_{\boldsymbol{z}}(P_1) - \mathsf{R}_{\boldsymbol{z}}(P_2) = \lambda\left(\mathsf{D}\left(P_1 \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}\left(P_2 \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}(P_2\|Q) - \mathsf{D}(P_1\|Q)\right).$$

Theorem (Theorem 8 in [Zou-2024])

For all $P_1 \in \triangle_{P_S}(\mathcal{X}\times\mathcal{Y})$ *and for all* $P_2 \in \triangle_{P_S}(\mathcal{X}\times\mathcal{Y})$,

$$\mathsf{R}_{\boldsymbol{\theta}}(P_1) - \mathsf{R}_{\boldsymbol{\theta}}(P_2) = \beta\left(\mathsf{D}\left(P_2 \| P^{(P_S,\beta)}_{\hat{\boldsymbol{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - \mathsf{D}\left(P_1 \| P^{(P_S,\beta)}_{\hat{\boldsymbol{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - \mathsf{D}(P_2\|P_S) + \mathsf{D}(P_1\|P_S)\right).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

[Zou-2024] Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman, and H. Vincent Poor. "**The Worst-Case Data-Generating Probability Measure in Statistical Learning**". IEEE Journal on Selected Areas in Information Theory, vol. 5, pp. 175–189, Apr., 2024.

Table of Contents

Empirical Risk **Optimization** with Relative Entropy Regularization

The Method of Gaps

Explicit Expressions for the Generalization Error

Concluding Remarks

Definition (Expected Empirical Risk)

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta})$$

$$\mathsf{R}_{\boldsymbol{\theta}}(Q) = \int \ell(h(\boldsymbol{\theta}, x), y) \, \mathrm{d}Q(x, y).$$

Definition (Expected Empirical Risk)

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta})$$

$$\mathsf{R}_{\boldsymbol{\theta}}(Q) = \int \ell(h(\boldsymbol{\theta}, x), y) \, \mathrm{d}Q(x, y).$$

Two **essential** observations:

- The generalization error is an expectation of the **variations of** $\mathsf{R}_{\boldsymbol{z}}$ **or** $\mathsf{R}_{\boldsymbol{\theta}}$; and

Definition (Expected Empirical Risk)

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta})$$

$$\mathsf{R}_{\boldsymbol{\theta}}(Q) = \int \ell(h(\boldsymbol{\theta}, x), y) \, \mathrm{d}Q(x, y).$$

Two **essential** observations:

- The generalization error is an expectation of the **variations of** $\mathsf{R}_{\boldsymbol{z}}$ **or** $\mathsf{R}_{\boldsymbol{\theta}}$; and
- These variations, a.k.a. **gaps**, exhibit closed-form expressions in terms of **information measures**.

Definition (Expected Empirical Risk)

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta})$$

$$\mathsf{R}_{\boldsymbol{\theta}}(Q) = \int \ell(h(\boldsymbol{\theta}, x), y) \, \mathrm{d}Q(x, y).$$

Two **essential** observations:

- The generalization error is an expectation of the **variations of** $\mathsf{R}_{\boldsymbol{z}}$ **or** $\mathsf{R}_{\boldsymbol{\theta}}$; and
- These variations, a.k.a. **gaps**, exhibit closed-form expressions in terms of **information measures**.

Two-step Method:

- To express the generalization error as an expectation of a gap; and

Definition (Expected Empirical Risk)

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta})$$

$$\mathsf{R}_{\boldsymbol{\theta}}(Q) = \int \ell(h(\boldsymbol{\theta}, x), y) \, \mathrm{d}Q(x, y).$$

Two **essential** observations:

- The generalization error is an expectation of the **variations of** $\mathsf{R}_{\boldsymbol{z}}$ **or** $\mathsf{R}_{\boldsymbol{\theta}}$; and
- These variations, a.k.a. **gaps**, exhibit closed-form expressions in terms of **information measures**.

Two-step Method:

- To express the generalization error as an expectation of a gap; and
- To leverage the properties of gaps to obtain closed-form expressions.

Definition (Expected Empirical Risk)

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta})$$

$$\mathsf{R}_{\boldsymbol{\theta}}(Q) = \int \ell(h(\boldsymbol{\theta}, x), y) \, \mathrm{d}Q(x, y).$$

Definition (Expected-Empirical-Risk **Gaps**)

Let functionals $\mathsf{G} : (\mathcal{X} \times \mathcal{Y})^n \times \triangle(\mathcal{M}) \times \triangle(\mathcal{M}) \to \mathbb{R}$ and $\mathsf{G} : \mathcal{M} \times \triangle(\mathcal{X} \times \mathcal{Y}) \times \triangle(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ be

$$\mathsf{G}(\boldsymbol{z}, P_1, P_2) = \mathsf{R}_{\boldsymbol{z}}(P_1) - \mathsf{R}_{\boldsymbol{z}}(P_2), \text{ **Algorithm-driven** Gap}$$

and

$$\mathsf{G}(\boldsymbol{\theta}, P_1, P_2) = \mathsf{R}_{\boldsymbol{\theta}}(P_1) - \mathsf{R}_{\boldsymbol{\theta}}(P_2). \text{ **Data-driven** Gap}$$

The Method of Gaps

Two variants:

- The Method of **Algorithm-driven** Gaps
  - Central building-block: **The Gibbs Algorithm**
  - No assumptions on $P_{\boldsymbol{Z}}$ (probability distribution of the datasets)
- The Method of **Data-driven** Gaps
  - Central building-block: **The Worst-Case Data-Generating** (WCDG) probability measure
  - I.I.D assumption on $P_{\boldsymbol{Z}}$:

$$P_{\boldsymbol{Z}}\left(\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n\right) = \prod_{t=1}^{n} P_Z\left(\mathcal{A}_t\right)$$

Generalization Error

The generalization error of the algorithm $P_{\Theta|Z}$ is

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) \triangleq \int \int \left(\mathrm{R}_{\boldsymbol{u}}\left(P_{\Theta|Z=z}\right) - \mathrm{R}_{\boldsymbol{z}}\left(P_{\Theta|Z=z}\right)\right) \mathrm{d}P_Z\left(\boldsymbol{u}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

Generalization Error

The generalization error of the algorithm $P_{\Theta|Z}$ is

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) \triangleq \int \int \left(\mathsf{R}_{\boldsymbol{u}}\left(P_{\Theta|Z=z}\right) - \mathsf{R}_{\boldsymbol{z}}\left(P_{\Theta|Z=z}\right)\right) \mathrm{d}P_Z\left(\boldsymbol{u}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

Step 1:

Lemma (Lemma 3 in [Perlaza-2024b])

The generalization error $\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = \int \mathsf{G}\left(\boldsymbol{z}, P_\Theta, P_{\Theta|Z=z}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right),$$

where for all measurable subsets $\mathcal{C}$ *of* $\mathcal{M}$*,*

$$P_\Theta\left(\mathcal{C}\right) = \int P_{\Theta|Z=z}\left(\mathcal{C}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Generalization Error

The generalization error of the algorithm $P_{\Theta|Z}$ is

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) \triangleq \int \int \left(\mathrm{R}_{\boldsymbol{u}}\left(P_{\Theta|Z=\boldsymbol{z}}\right) - \mathrm{R}_{\boldsymbol{z}}\left(P_{\Theta|Z=\boldsymbol{z}}\right)\right) \mathrm{d}P_Z\left(\boldsymbol{u}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

Step 2:

Generalization Error

The generalization error of the algorithm $P_{\Theta|Z}$ is

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) \triangleq \int \int \left(\mathrm{R}_{\boldsymbol{u}}\left(P_{\Theta|Z=z}\right) - \mathrm{R}_z\left(P_{\Theta|Z=z}\right)\right) \mathrm{d}P_Z\left(\boldsymbol{u}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

Step 2:

Lemma (Lemma 4 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = \lambda \int \left(\mathrm{D}\left(P_\Theta \| P^{(Q,\lambda)}_{\Theta|Z=z}\right) - \mathrm{D}\left(P_{\Theta|Z=z} \| P^{(Q,\lambda)}_{\Theta|Z=z}\right) + \mathrm{D}\left(P_{\Theta|Z=z} \| Q\right) - \mathrm{D}\left(P_\Theta \| Q\right)\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Generalization Error

The generalization error of the algorithm $P_{\Theta|Z}$ is

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) \triangleq \int \int \left(\mathsf{R}_{\boldsymbol{u}}\left(P_{\Theta|Z=z}\right) - \mathsf{R}_{\boldsymbol{z}}\left(P_{\Theta|Z=z}\right)\right) \mathrm{d}P_Z\left(\boldsymbol{u}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

Step 1:

- **Assumption:** $P_Z\left(\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n\right) = \prod_{t=1}^{n} P_Z\left(\mathcal{A}_t\right)$.

Generalization Error

The generalization error of the algorithm $P_{\Theta|Z}$ is

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) \triangleq \int \int \left(\mathsf{R}_{\boldsymbol{u}}\left(P_{\Theta|Z=z}\right) - \mathsf{R}_{\boldsymbol{z}}\left(P_{\Theta|Z=z}\right)\right) \mathrm{d}P_Z\left(\boldsymbol{u}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

Step 1:

- **Assumption:** $P_Z\left(\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n\right) = \prod_{t=1}^{n} P_Z\left(\mathcal{A}_t\right)$.

Lemma (Lemma 6 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_Z) = \int \mathsf{G}\left(\boldsymbol{\theta}, P_Z, P_{Z|\Theta=\boldsymbol{\theta}}\right) \mathrm{d}P_\Theta\left(\boldsymbol{\theta}\right).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Generalization Error

The generalization error of the algorithm $P_{\Theta|\boldsymbol{Z}}$ is

$$\overline{\overline{G}}\left(P_{\Theta|\boldsymbol{Z}}, P_{\boldsymbol{Z}}\right) \triangleq \int \int \left(\mathrm{R}_{\boldsymbol{u}}\left(P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathrm{R}_{\boldsymbol{z}}\left(P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right)\right) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{u}\right) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{z}\right).$$

Step 2:

- **Assumption:** $P_{\boldsymbol{Z}}\left(\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n\right) = \prod_{t=1}^{n} P_Z\left(\mathcal{A}_t\right)$.

Generalization Error

The generalization error of the algorithm $P_{\Theta|Z}$ is

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) \triangleq \int \int \left(\mathrm{R}_{\boldsymbol{u}}\left(P_{\Theta|Z=\boldsymbol{z}}\right) - \mathrm{R}_{\boldsymbol{z}}\left(P_{\Theta|Z=\boldsymbol{z}}\right)\right) \mathrm{d}P_Z\left(\boldsymbol{u}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

Step 2:

- **Assumption:** $P_Z\left(\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n\right) = \prod_{t=1}^{n} P_Z\left(\mathcal{A}_t\right)$.

Lemma (Lemma 7 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z) = \beta \int \left(\mathrm{D}\left(P_{Z|\Theta=\boldsymbol{\theta}} \| P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}\right) - \mathrm{D}\left(P_Z \| P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}\right) - \mathrm{D}\left(P_{Z|\Theta=\boldsymbol{\theta}} \| P_S\right) + \mathrm{D}\left(P_Z \| P_S\right) \right) \mathrm{d}P_\Theta\left(\boldsymbol{\theta}\right).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

So far...

Generalization Error

The generalization error of the algorithm $P_{\Theta|Z}$ is

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) \triangleq \int \int \left(\mathrm{R}_{\boldsymbol{u}}\left(P_{\Theta|Z=z}\right) - \mathrm{R}_{\boldsymbol{z}}\left(P_{\Theta|Z=z}\right)\right) \mathrm{d}P_Z\left(\boldsymbol{u}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

So far...

Generalization Error

The generalization error of the algorithm $P_{\Theta|Z}$ is

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) \triangleq \int \int \left(\mathrm{R}_{\boldsymbol{u}}\left(P_{\Theta|Z=z}\right) - \mathrm{R}_{\boldsymbol{z}}\left(P_{\Theta|Z=z}\right)\right) \mathrm{d}P_Z\left(\boldsymbol{u}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

Lemma (Lemma 4 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = \lambda \int \left(\mathrm{D}\left(P_\Theta \| P^{(Q,\lambda)}_{\Theta|Z=z}\right) - \mathrm{D}\left(P_{\Theta|Z=z} \| P^{(Q,\lambda)}_{\Theta|Z=z}\right) + \mathrm{D}\left(P_{\Theta|Z=z} \| Q\right) - \mathrm{D}\left(P_\Theta \| Q\right)\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

So far...

Generalization Error

The generalization error of the algorithm $P_{\Theta|Z}$ is

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) \triangleq \int \int \left(\mathrm{R}_{\boldsymbol{u}}\left(P_{\Theta|Z=z}\right) - \mathrm{R}_{\boldsymbol{z}}\left(P_{\Theta|Z=z}\right)\right) \mathrm{d}P_Z\left(\boldsymbol{u}\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

Lemma (Lemma 4 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = \lambda \int \left(\mathrm{D}\left(P_\Theta \| P^{(Q,\lambda)}_{\Theta|Z=z}\right) - \mathrm{D}\left(P_{\Theta|Z=z} \| P^{(Q,\lambda)}_{\Theta|Z=z}\right) + \mathrm{D}\left(P_{\Theta|Z=z} \| Q\right) - \mathrm{D}\left(P_\Theta \| Q\right)\right) \mathrm{d}P_Z\left(\boldsymbol{z}\right).$$

Lemma (Lemma 7 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z) = \beta \int \left(\mathrm{D}\left(P_{Z|\Theta=\boldsymbol{\theta}} \| P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}\right) - \mathrm{D}\left(P_Z \| P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}\right) - \mathrm{D}\left(P_{Z|\Theta=\boldsymbol{\theta}} \| P_S\right) + \mathrm{D}\left(P_Z \| P_S\right)\right) \mathrm{d}P_\Theta\left(\boldsymbol{\theta}\right).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Table of Contents

Empirical Risk **Optimization** with Relative Entropy Regularization

The Method of Gaps

Explicit Expressions for the Generalization Error

Concluding Remarks

Expressions Obtained Via the Method of Gaps

- Particular **choices of the parameters**

Expressions Obtained Via the Method of Gaps

- Particular **choices of the parameters**
- **Algebraic manipulations** of the closed-form expressions shown before

Expressions Obtained Via the Method of Gaps

- Particular **choices of the parameters**
- **Algebraic manipulations** of the closed-form expressions shown before
- **More manipulations lead to less generality**
- Additional conditions to allow manipulations are imposed on:
  - The algorithm; and
  - The data-generating distribution.

Expressions Obtained Via the Method of Gaps

- Particular **choices of the parameters**
- **Algebraic manipulations** of the closed-form expressions shown before
- **More manipulations lead to less generality**
- Additional conditions to allow manipulations are imposed on:
  - The algorithm; and
  - The data-generating distribution.
- Some Expressions establish **bridges with other areas**: Hypothesis Testing, Geometry, etc.

Theorem (Theorem 8 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\begin{aligned}&\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right)\\ &= \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta})\, P_Z(\boldsymbol{z}) - \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta}(\boldsymbol{\theta})\, \mathrm{d}P_Z(\boldsymbol{z}).\end{aligned}$$

Expressions Obtained Via the Method of Gaps

Connections to Hypothesis Testing

Theorem (Theorem 8 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\begin{aligned}&\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right)\\ &= \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta})\, P_Z(\boldsymbol{z}) - \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta}(\boldsymbol{\theta})\, \mathrm{d}P_Z(\boldsymbol{z}).\end{aligned}$$

Statistical Hypothesis Test

- Ground truth probability distribution: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim P_{\Theta|Z} \cdot P_Z$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Expressions Obtained Via the Method of Gaps

Connections to Hypothesis Testing

Theorem (Theorem 8 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta})\, P_Z(\boldsymbol{z}) - \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\Theta}(\boldsymbol{\theta})\, \mathrm{d}P_Z(\boldsymbol{z}).$$

Statistical Hypothesis Test

- Ground truth probability distribution: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim P_{\Theta|Z} \cdot P_Z$
- Null Hypothesis: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim P^{(Q,\lambda)}_{\Theta|Z} \cdot P_Z$
- Alternative Hypothesis: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim \cdot P_Z$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Theorem (Theorem 8 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta})\, P_Z(\boldsymbol{z}) - \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta}(\boldsymbol{\theta})\, \mathrm{d}P_Z(\boldsymbol{z}).$$

Statistical Hypothesis Test

- Ground truth probability distribution: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim P_{\Theta|Z} \cdot P_Z$
- Null Hypothesis: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim P^{(Q,\lambda)}_{\Theta|Z} \cdot P_Z$
- Alternative Hypothesis: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim \cdot P_Z$
- log-likelihood ratio:

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z} \cdot P_Z}{\mathrm{d}Q \cdot P_Z}(\boldsymbol{\theta}, \boldsymbol{z}) = \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Expressions Obtained Via the Method of Gaps

Connections to Hypothesis Testing

Theorem (Theorem 8 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta})\, P_Z(\boldsymbol{z}) - \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta}(\boldsymbol{\theta})\, \mathrm{d}P_Z(\boldsymbol{z}).$$

Statistical Hypothesis Test

- Ground truth probability distribution: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim P_{\Theta|Z} \cdot P_Z$
- Null Hypothesis: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim P^{(Q,\lambda)}_{\Theta|Z} \cdot P_Z$
- Alternative Hypothesis: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim \mathrm{d}Q \cdot P_Z$
- **Mismatched** Hypothesis Test [Boroumand-2022]

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

[Boroumand-2022] P. Boroumand and A. G. i Fàbregas, "**Mismatched binary hypothesis testing: Error exponent sensitivity**," IEEE Transactions on Information Theory, vol. 68, no. 10, pp. 6738 – 6761, 2022.

Theorem (Theorem 8 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\begin{aligned}&\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right)\\ &= \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta})\, P_Z(\boldsymbol{z}) - \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta}(\boldsymbol{\theta})\, \mathrm{d}P_Z(\boldsymbol{z}).\end{aligned}$$

Statistical Hypothesis Test

- Ground truth probability distribution: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim P_{\Theta|Z} \cdot P_Z$
- Null Hypothesis: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim P^{(Q,\lambda)}_{\Theta|Z} \cdot P_Z$
- Alternative Hypothesis: $(\boldsymbol{\Theta}, \boldsymbol{Z}) \sim \mathrm{d}Q \cdot P_Z$
- **Mismatched** Hypothesis Test [Boroumand-2022]

Generalization Error

variation of the expected log-likelihood when the ground-truth changes from $P_\Theta \cdot P_Z$ to $P_{\Theta|Z} \cdot P_Z$.

Expressions Obtained Via the Method of Gaps

Connections to Hypothesis Testing

Theorem (Theorem 8 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta})\, P_Z(z) - \lambda \int\int \left(\log \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P_{\Theta}(\boldsymbol{\theta})\, \mathrm{d}P_Z(z).$$

Theorem (Theorem 22 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_Z) = \beta\left(\int\int \log\left(\frac{\mathrm{d}P_S}{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}}(z)\right)\mathrm{d}P_{Z|\Theta=\boldsymbol{\theta}}(z)\mathrm{d}P_{\Theta}(\boldsymbol{\theta}) - \int\int \log\left(\frac{\mathrm{d}P_S}{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}}(z)\right)\mathrm{d}P_Z(z)\mathrm{d}P_{\Theta}(\boldsymbol{\theta})\right).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Corollary (Corollary 9 in [Perlaza-2024b] – **Choice of** $Q = P_{\Theta}$ *)*

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = \lambda I\left(P_{\Theta|Z}; P_Z\right) + \lambda \int \left(\mathrm{D}\left(P_{\Theta} \| P_{\Theta|Z=z}^{(P_{\Theta},\lambda)}\right) - \mathrm{D}\left(P_{\Theta|Z=z} \| P_{\Theta|Z=z}^{(P_{\Theta},\lambda)}\right) \right) \mathrm{d}P_Z\left(z\right).$$

Corollary (Corollary 9 in [Perlaza-2024b] – Choice of $Q = P_\Theta$)

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = \lambda I\left(P_{\Theta|Z}; P_Z\right) + \lambda \int \left(\mathrm{D}\left(P_\Theta \| P^{(P_\Theta,\lambda)}_{\Theta|Z=z}\right) - \mathrm{D}\left(P_{\Theta|Z=z} \| P^{(P_\Theta,\lambda)}_{\Theta|Z=z}\right) \right) \mathrm{d}P_Z(z).$$

Corollary (Corollary 24 in [Perlaza-2024b] – Choice of $P_S = P_Z$)

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\begin{aligned}&\overline{\overline{G}}(P_{\Theta|Z}, P_Z)\\ =&-\beta I\left(P_{Z|\Theta}; P_\Theta\right) + \beta \int \mathrm{D}\left(P_{Z|\Theta=\theta} \| P^{(P_Z,\beta)}_{\hat{Z}|\Theta=\theta}\right) \mathrm{d}P_\Theta(\theta) - \beta \int \mathrm{D}\left(P_Z \| P^{(P_Z,\beta)}_{\hat{Z}|\Theta=\theta}\right) \mathrm{d}P_\Theta(\theta).\end{aligned}$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Corollary (Corollary 10 in [Perlaza-2024b] – **Choice of** $Q = P_{\Theta|Z}$*)*

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\begin{aligned}&\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right)\\ &= -\lambda L\left(P_{\Theta|Z}; P_Z\right) + \lambda \int \mathrm{D}\left(P_\Theta \| P_{\Theta|Z=z}^{\left(P_{\Theta|Z=z'}, \lambda\right)}\right) \mathrm{d}P_Z(z) - \lambda \int \mathrm{D}\left(P_{\Theta|Z=z} \| P_{\Theta|Z=z}^{\left(P_{\Theta|Z=z'}, \lambda\right)}\right) \mathrm{d}P_Z(z).\end{aligned}$$

Corollary (Corollary 10 in [Perlaza-2024b] – Choice of $Q = P_{\Theta|Z}$)

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}\left(P_{\Theta|Z}, P_Z\right) = -\lambda L\left(P_{\Theta|Z}; P_Z\right) + \lambda \int \mathrm{D}\left(P_\Theta \| P_{\Theta|Z=z}^{\left(P_{\Theta|Z=z},\lambda\right)}\right) \mathrm{d}P_Z(z) - \lambda \int \mathrm{D}\left(P_{\Theta|Z=z} \| P_{\Theta|Z=z}^{\left(P_{\Theta|Z=z},\lambda\right)}\right) \mathrm{d}P_Z(z).$$

Corollary (Corollary 25 in [Perlaza-2024b] – Choice of $P_S = P_{Z|\Theta}$)

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}(P_{\Theta|Z}, P_Z) = \beta L\left(P_{Z|\Theta}; P_\Theta\right) + \beta \int \mathrm{D}\left(P_{Z|\Theta=\boldsymbol{\theta}} \| P_{\hat{Z}|\Theta=\boldsymbol{\theta}}^{\left(P_{Z|\Theta=\boldsymbol{\theta}},\beta\right)}\right) \mathrm{d}P_\Theta(\boldsymbol{\theta}) - \beta \int \mathrm{D}\left(P_Z \| P_{\hat{Z}|\Theta=\boldsymbol{\theta}}^{\left(P_{Z|\Theta=\boldsymbol{\theta}},\beta\right)}\right) \mathrm{d}P_\Theta(\boldsymbol{\theta}).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Theorem (Theorem 14 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}(P_{\Theta|Z}, P_Z) = \lambda \left(I\left(P_{\Theta|Z}; P_Z\right) + L\left(P_{\Theta|Z}; P_Z\right)\right) + \lambda \int\int \log \frac{\mathrm{d}P_{\Theta|Z=z}}{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}(\boldsymbol{\theta}) \mathrm{d}P_{\Theta}(\boldsymbol{\theta}) \mathrm{d}P_{Z}(\boldsymbol{z}) - \lambda \int\int \log \frac{\mathrm{d}P_{\Theta|Z=z}}{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}(\boldsymbol{\theta}) \mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta}) \mathrm{d}P_{Z}(\boldsymbol{z}).$$

Theorem (Theorem 14 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}(P_{\Theta|Z}, P_Z) = \lambda \left(I\left(P_{\Theta|Z}; P_Z\right) + L\left(P_{\Theta|Z}; P_Z\right)\right) + \lambda \int\int \log \frac{\mathrm{d}P_{\Theta|Z=z}}{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}(\boldsymbol{\theta})\mathrm{d}P_{\Theta}(\boldsymbol{\theta})\mathrm{d}P_{Z}(\boldsymbol{z}) - \lambda \int\int \log \frac{\mathrm{d}P_{\Theta|Z=z}}{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}(\boldsymbol{\theta})\mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta})\mathrm{d}P_{Z}(\boldsymbol{z}).$$

What if...

$$\lambda \int\int \log \frac{\mathrm{d}P_{\Theta|Z=z}}{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}(\boldsymbol{\theta})\,\mathrm{d}P_{\Theta}(\boldsymbol{\theta})\,\mathrm{d}P_{Z}(\boldsymbol{z}) - \lambda \int\int \log \frac{\mathrm{d}P_{\Theta|Z=z}}{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}(\boldsymbol{\theta})\,\mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta})\,\mathrm{d}P_{Z}(\boldsymbol{z}) = 0.$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Theorem (Theorem 14 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}(P_{\Theta|Z}, P_Z) = \lambda \left(I\left(P_{\Theta|Z}; P_Z\right) + L\left(P_{\Theta|Z}; P_Z\right)\right) + \lambda \int\int \log \frac{\mathrm{d}P_{\Theta|Z=z}}{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}(\boldsymbol{\theta})\mathrm{d}P_{\Theta}(\boldsymbol{\theta})\mathrm{d}P_Z(\boldsymbol{z}) - \lambda \int\int \log \frac{\mathrm{d}P_{\Theta|Z=z}}{\mathrm{d}P^{(Q,\lambda)}_{\Theta|Z=z}}(\boldsymbol{\theta})\mathrm{d}P_{\Theta|Z=z}(\boldsymbol{\theta})\mathrm{d}P_Z(\boldsymbol{z}).$$

Generalization Error of the **Gibbs Algorithm**:

Corollary (Theorem 1 in [Aminian-2021])

$$\overline{\overline{G}}(P^{(Q,\lambda)}_{\Theta|Z}, P_Z) = \lambda \left(I\left(P^{(Q,\lambda)}_{\Theta|Z}; P_Z\right) + L\left(P^{(Q,\lambda)}_{\Theta|Z}; P_Z\right)\right).$$

[Aminian-2021] G Aminian, Y Bu, L Toni, M Rodrigues, G Wornell. "An exact characterization of the generalization error for the Gibbs algorithm" Advances in Neural Information Processing Systems, vol. 34, pp. 8106-8118, 2021

Theorem (Theorem 29 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_Z) = -\beta\big(I\left(P_{Z|\Theta}; P_\Theta\right) + L\left(P_{Z|\Theta}; P_\Theta\right)\big)$$
$$+\beta \int\int \log\left(\frac{\mathrm{d}P_{Z|\Theta=\boldsymbol{\theta}}}{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}}(z)\right)\mathrm{d}P_{Z|\Theta=\boldsymbol{\theta}}(z)\,\mathrm{d}P_\Theta(\boldsymbol{\theta}) - \beta\int\int \log\left(\frac{\mathrm{d}P_{Z|\Theta=\boldsymbol{\theta}}}{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}}(z)\right)\mathrm{d}P_Z(z)\,\mathrm{d}P_\Theta(\boldsymbol{\theta}).$$

Theorem (Theorem 29 in [Perlaza-2024b])

The generalization error $\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_Z) = -\beta\left(I\left(P_{Z|\Theta}; P_\Theta\right) + L\left(P_{Z|\Theta}; P_\Theta\right)\right)$$
$$+\beta \int\int \log\left(\frac{\mathrm{d}P_{Z|\Theta=\boldsymbol{\theta}}}{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}}(z)\right)\mathrm{d}P_{Z|\Theta=\boldsymbol{\theta}}(z)\,\mathrm{d}P_\Theta(\boldsymbol{\theta}) - \beta \int\int \log\left(\frac{\mathrm{d}P_{Z|\Theta=\boldsymbol{\theta}}}{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}}(z)\right)\mathrm{d}P_Z(z)\,\mathrm{d}P_\Theta(\boldsymbol{\theta}).$$

What if...

$$\beta \int\int \log\left(\frac{\mathrm{d}P_{Z|\Theta=\boldsymbol{\theta}}}{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}}(z)\right)\mathrm{d}P_{Z|\Theta=\boldsymbol{\theta}}(z)\,\mathrm{d}P_\Theta(\boldsymbol{\theta}) - \beta \int\int \log\left(\frac{\mathrm{d}P_{Z|\Theta=\boldsymbol{\theta}}}{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}}(z)\right)\mathrm{d}P_Z(z)\,\mathrm{d}P_\Theta(\boldsymbol{\theta}) = 0.$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Theorem (Theorem 18 in [Perlaza-2024b])

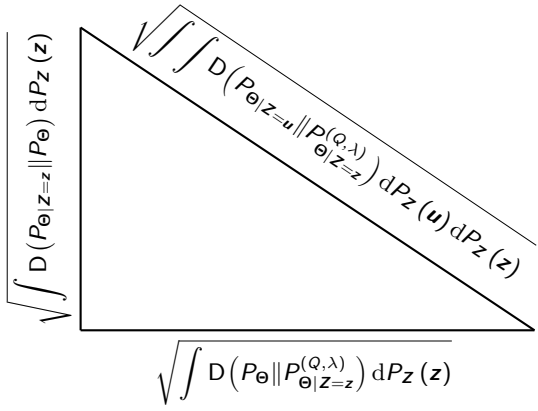
The generalization error $\overline{\overline{\mathrm{G}}}(P_{\Theta|Z}, P_Z)$ *satisfies*

$$\overline{\overline{G}}(P_{\Theta|Z}, P_Z) = \lambda \int \int \left(\mathrm{D}\left(P_{\Theta|Z=z} \| P^{(Q,\lambda)}_{\Theta|Z=u} \right) - \mathrm{D}\left(P_{\Theta|Z=z} \| P^{(Q,\lambda)}_{\Theta|Z=z} \right) \right) \mathrm{d}P_Z(u)\, \mathrm{d}P_Z(z).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Expressions Obtained Via the Method of Gaps

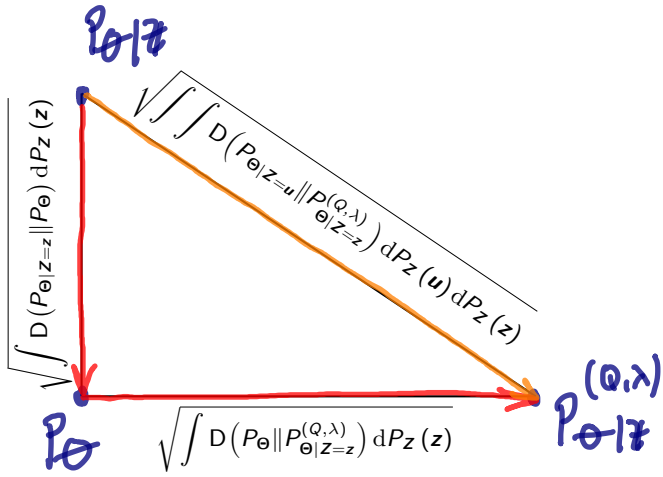
Connections to Euclidian Geometry



$$\int\int \mathrm{D}\left(P_{\Theta|Z=u}\|P^{(Q,\lambda)}_{\Theta|Z=z}\right)\mathrm{d}P_{Z}(u)\,\mathrm{d}P_{Z}(z) = \int \mathrm{D}\left(P_{\Theta}\|P^{(Q,\lambda)}_{\Theta|Z=z}\right)\mathrm{d}P_{Z}(z) + \int \mathrm{D}\left(P_{\Theta|Z=z}\|P_{\Theta}\right)\mathrm{d}P_{Z}(z)\,.$$

Expressions Obtained Via the Method of Gaps

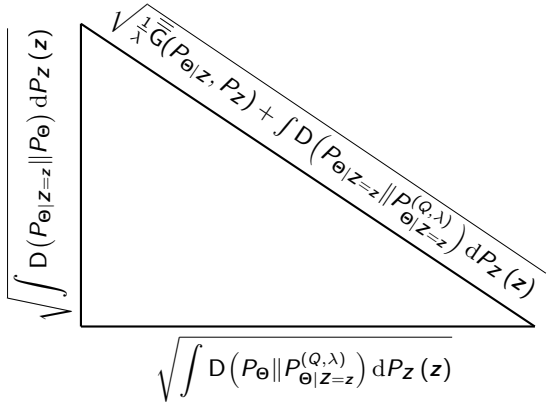
Connections to Euclidian Geometry



$$\int\int \mathrm{D}\left(P_{\Theta|Z=u} \| P^{(Q,\lambda)}_{\Theta|Z=z}\right) \mathrm{d}P_Z(u)\,\mathrm{d}P_Z(z) = \int \mathrm{D}\left(P_\Theta \| P^{(Q,\lambda)}_{\Theta|Z=z}\right) \mathrm{d}P_Z(z) + \int \mathrm{D}\left(P_{\Theta|Z=z} \| P_\Theta\right) \mathrm{d}P_Z(z).$$

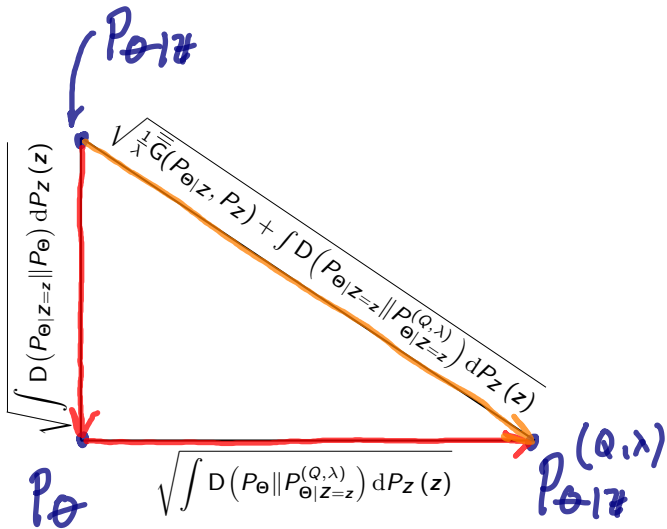
Expressions Obtained Via the Method of Gaps

Connections to Euclidian Geometry



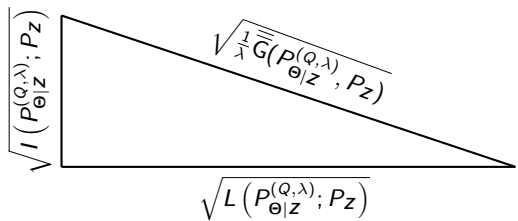
Expressions Obtained Via the Method of Gaps

Connections to Euclidian Geometry



Expressions Obtained Via the Method of Gaps

Connections to Euclidian Geometry



Theorem (Theorem 31 in [Perlaza-2024b])

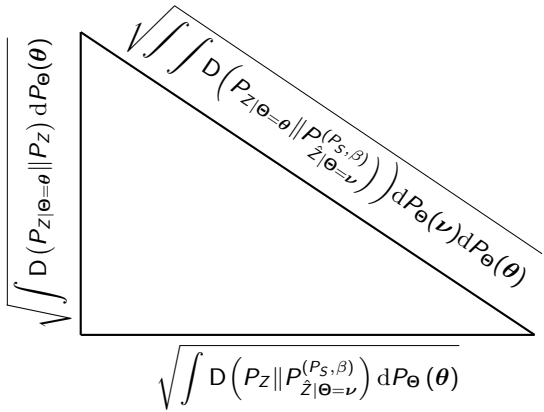
The generalization error $\overline{\overline{\mathsf{G}}}(P_{\mathbf{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ *satisfies*

$$\overline{\overline{\mathsf{G}}}\left(P_{\mathbf{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}\right) = \beta \int \int \left(\mathsf{D}\left(P_{Z|\mathbf{\Theta}=\boldsymbol{\mu}} \| P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\mu}} \right) - \mathsf{D}\left(P_{Z|\mathbf{\Theta}=\boldsymbol{\mu}} \| P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\nu}} \right) \right) \mathrm{d}P_{\mathbf{\Theta}}(\boldsymbol{\nu}) \mathrm{d}P_{\mathbf{\Theta}}(\boldsymbol{\mu}).$$

[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.

Expressions Obtained Via the Method of Gaps

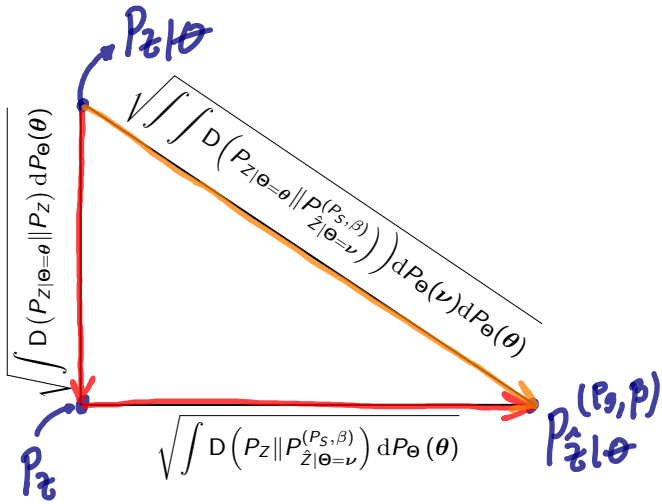
Connections to Euclidian Geometry



$$\int\int \mathrm{D}\left(P_{Z|\Theta=\boldsymbol{\theta}} \| P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\nu}}\right) \mathrm{d}P_\Theta(\boldsymbol{\nu}) \mathrm{d}P_\Theta(\boldsymbol{\theta}) = \int \mathrm{D}\left(P_{Z|\Theta=\boldsymbol{\theta}} \| P_Z\right) \mathrm{d}P_\Theta(\boldsymbol{\theta}) + \int \mathrm{D}\left(P_Z \| P^{(P_S,\beta)}_{\hat{Z}|\Theta=\boldsymbol{\theta}}\right) \mathrm{d}P_\Theta(\boldsymbol{\theta}),$$

Expressions Obtained Via the Method of Gaps

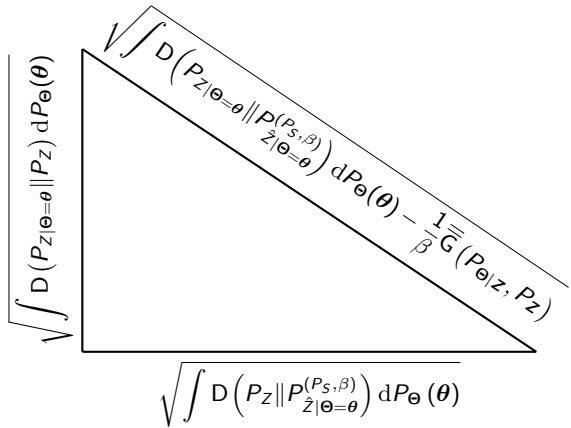
Connections to Euclidian Geometry



$$\int\int \mathrm{D}\left(P_{Z|\Theta=\theta}\|P^{(P_S,\beta)}_{\hat{Z}|\Theta=\nu}\right)\mathrm{d}P_\Theta(\nu)\mathrm{d}P_\Theta(\theta) = \int \mathrm{D}\left(P_{Z|\Theta=\theta}\|P_Z\right)\mathrm{d}P_\Theta(\theta) + \int \mathrm{D}\left(P_Z\|P^{(P_S,\beta)}_{\hat{Z}|\Theta=\theta}\right)\mathrm{d}P_\Theta(\theta),$$

Expressions Obtained Via the Method of Gaps

Connections to Euclidian Geometry



Expressions Obtained Via the Method of Gaps

Connections to Euclidian Geometry

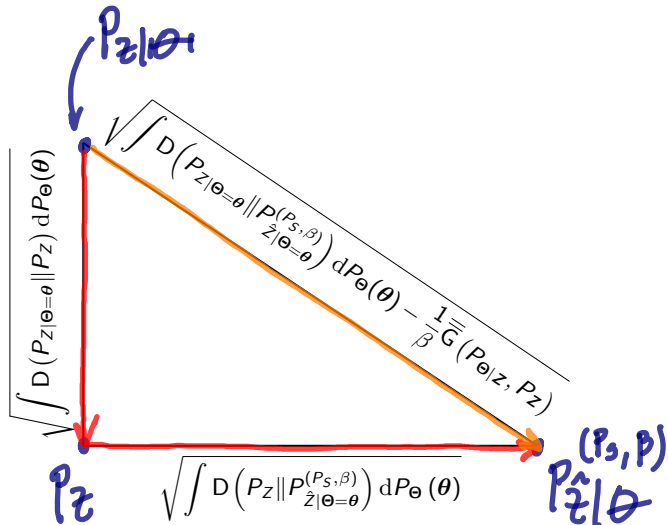


Table of Contents

Empirical Risk **Optimization** with Relative Entropy Regularization

The Method of Gaps

Explicit Expressions for the Generalization Error

Concluding Remarks

Some Concluding Remarks

- Solution to the **Empirical Risk Optimization with Relative Entropy regularization**
  - Maximization $\rightarrow$ **Worst-Case Data-Generating Probability Measure**
  - Minimization $\rightarrow$ **Gibbs Algorithm**

Some Concluding Remarks

- Solution to the **Empirical Risk Optimization with Relative Entropy regularization**
  - Maximization $\rightarrow$ **Worst-Case Data-Generating Probability Measure**
  - Minimization $\rightarrow$ **Gibbs Algorithm**
- Method of Gaps
  - **Algorithm-driven** gaps $\rightarrow$ uses **Gibbs Algorithm**
  - **Data-driven** gaps $\rightarrow$ uses **Worst-Case Data-Generating Probability Measure**

Some Concluding Remarks

- Solution to the **Empirical Risk Optimization with Relative Entropy regularization**
  - Maximization $\rightarrow$ **Worst-Case Data-Generating Probability Measure**
  - Minimization $\rightarrow$ **Gibbs Algorithm**
- Method of Gaps
  - **Algorithm-driven** gaps $\rightarrow$ uses **Gibbs Algorithm**
  - **Data-driven** gaps $\rightarrow$ uses **Worst-Case Data-Generating Probability Measure**
- **Generalization error** obtained via
  - Expectation of **Algorithm-driven** gaps
  - Expectation of **Data-driven** gaps

WHAT IS THE LONG-RUN DISTRIBUTION OF STOCHASTIC GRADIENT DESCENT? A LARGE DEVIATIONS ANALYSIS

WAÏSS AZIZIAN[c,*], FRANCK IUTZELER[♯],
JÉRÔME MALICK*, AND PANAYOTIS MERTIKOPOULOS[⋄]

Abstract. In this paper, we examine the long-run distribution of stochastic gradient descent (SGD) in general, non-convex problems. Specifically, we seek to understand which regions of the problem's state space are more likely to be visited by SGD, and by how much. Using an approach based on the theory of large deviations and randomly perturbed dynamical systems, we show that the long-run distribution of SGD resembles the Boltzmann–Gibbs distribution of equilibrium thermodynamics with temperature equal to the method's step-size and energy levels determined by the problem's objective and the statistics of the noise. In particular, we show that, in the long run, (*a*) the problem's critical region is visited exponentially more often than any non-critical region;

Examples

Corollary (What is the long-run Generalization Error of Stochastic Gradient Descent ?)

$$\overline{\overline{G}}(P_{\Theta|Z}^{(Q,\lambda)}, P_Z) = \lambda \left(I\left(P_{\Theta|Z}^{(Q,\lambda)}; P_Z\right) + L\left(P_{\Theta|Z}^{(Q,\lambda)}; P_Z\right)\right).$$

Thank you for your attention!

Questions/Comments/Typos: samir.perlaza@inria.fr

- This work appears in:
  - Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.
  - Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, and Stefano Rini. "**Empirical Risk Minimization with Relative Entropy Regularization**". IEEE Transactions on Information Theory, vol. 70, no. 7, pp. 5122 – 5161, July, 2024.
  - Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman, and H. Vincent Poor. "**The Worst-Case Data-Generating Probability Measure in Statistical Learning**". IEEE Journal on Selected Areas in Information Theory, vol. 5, pp. 175–189, Apr., 2024.