Tutorial

Characterizing the Generalization Error of Machine Learning Algorithms via Information Measures

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

2024 IEEE Information Theory Workshop

The 24th of November, 2024
Shenzhen, China

Slides for Part III

Table of Contents

Information Source

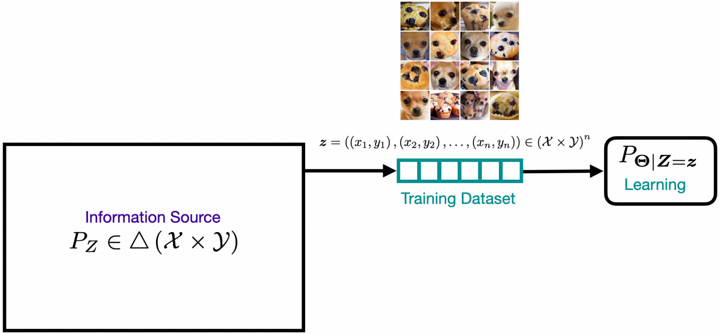
$$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$$

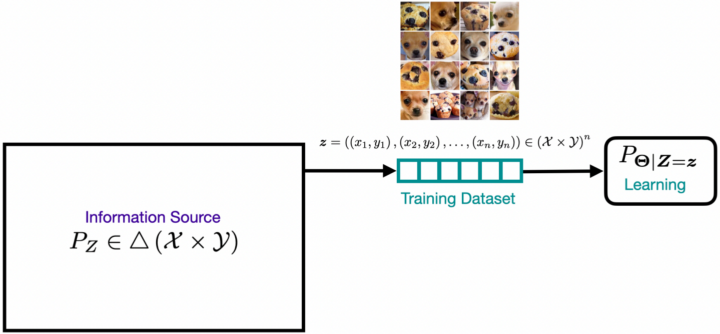
Information Source

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

Information Source

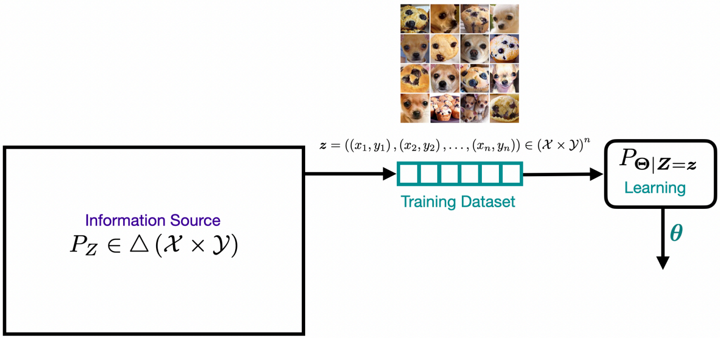
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$


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

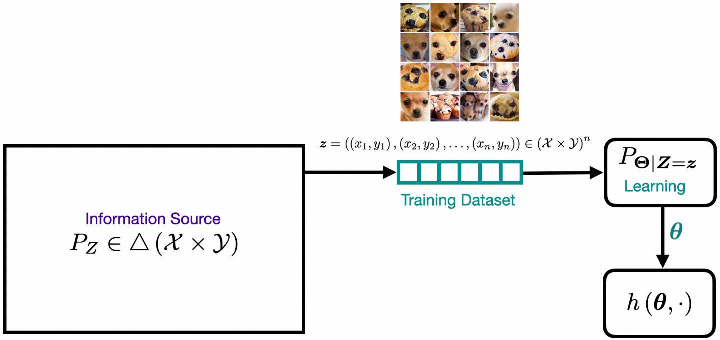


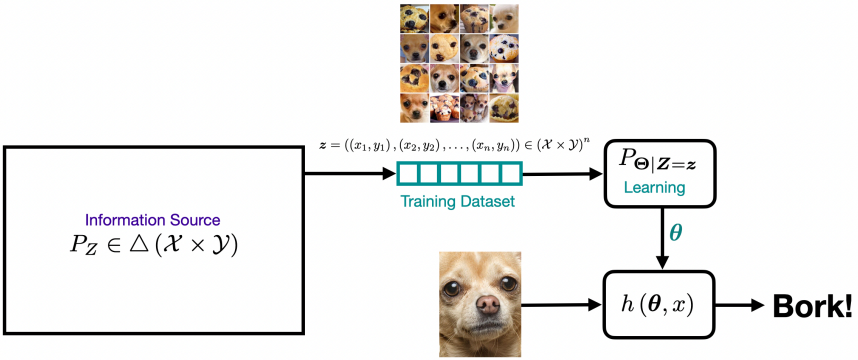
Algorithm

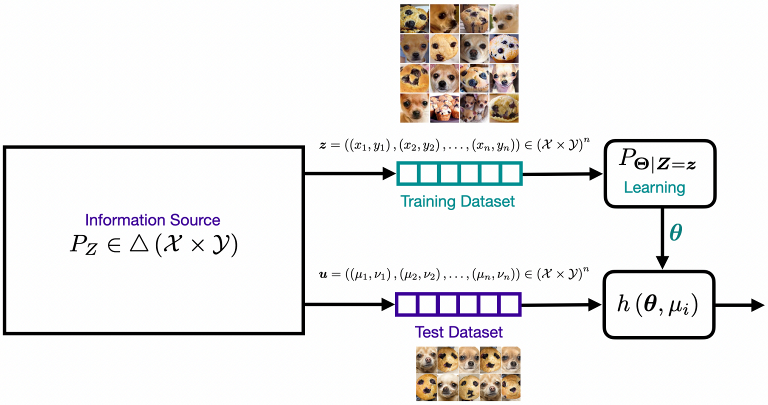
A conditional probability measure $P_{\Theta|Z} \in \triangle\left(\mathcal{M} | (\mathcal{X} \times \mathcal{Y})^n\right)$ represents a supervised machine learning algorithm.



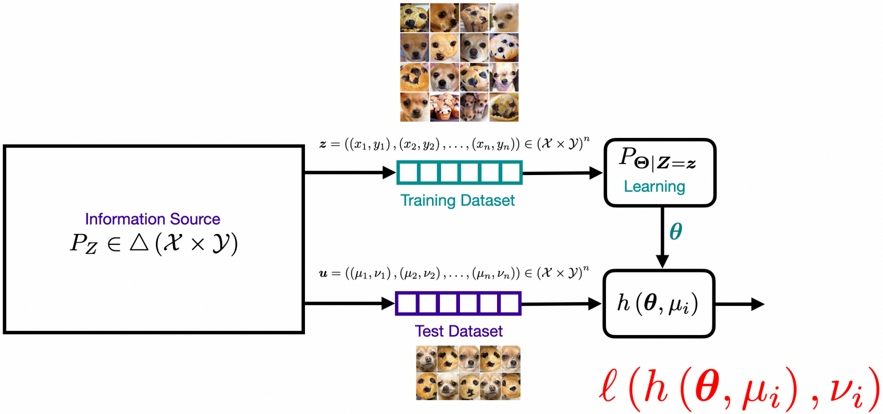
$z = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Training Dataset
$P_{\Theta | Z = z}$
Learning
θ
Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$

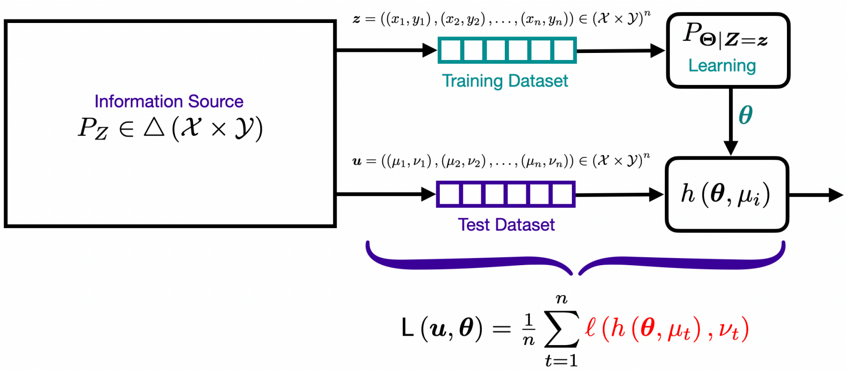

Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
$\boldsymbol{z} = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Training Dataset
$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$
Learning
$\boldsymbol{\theta}$
$h(\boldsymbol{\theta},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

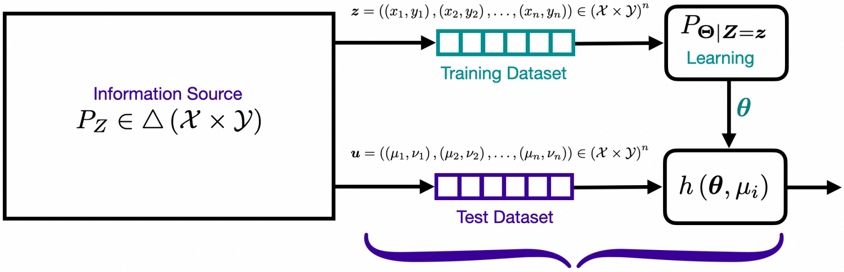

$\boldsymbol{z} = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
Information Source
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
Training Dataset
$P_{\Theta | \boldsymbol{Z} = \boldsymbol{z}}$
Learning
$\boldsymbol{\theta}$
$h(\boldsymbol{\theta}, x)$
Bork!



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
$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$
$\boldsymbol{z} = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$
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
$h(\boldsymbol{\theta}, \mu_i)$
$\ell(h(\boldsymbol{\theta}, \mu_i), \nu_i)$


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
$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$
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
$h(\boldsymbol{\theta}, \mu_i)$
$\mathsf{L}(\boldsymbol{u}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^{n} \ell(h(\boldsymbol{\theta}, \mu_t), \nu_t)$



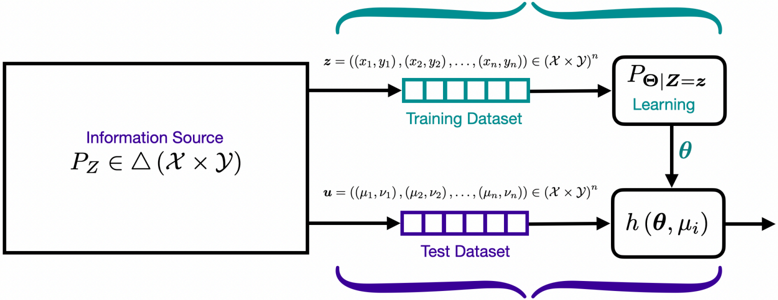
$$\mathsf{L}(\boldsymbol{u}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^{n} \ell(h(\boldsymbol{\theta}, \mu_t), \nu_t)$$

Problem Formulation: Empirical Risk Minimization (ERM)

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

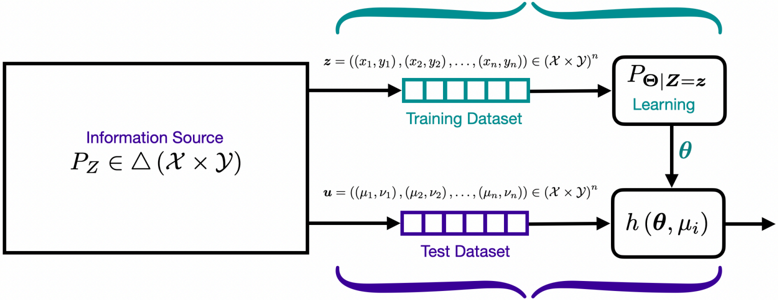
$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}).$$

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



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

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



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

Training (Expected) Risk and Test (Expected) Risk

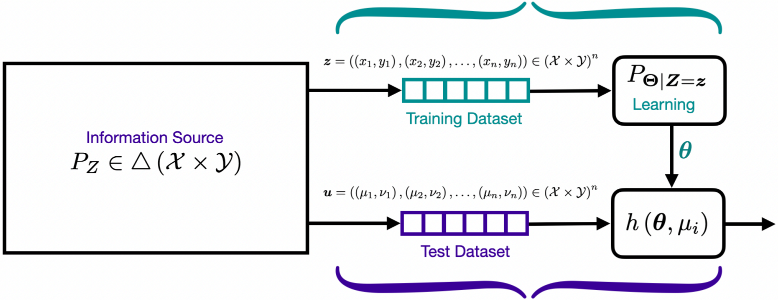
$$\underbrace{\mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}_{\text{Test Expected Risk}} - \underbrace{\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}_{\text{Training Expected Risk}}$$

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:

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

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 $Q \in \triangle\left(\mathcal{M}, \mathscr{B}\left(\mathcal{M}\right)\right)$ and $\lambda \in (0, +\infty)$, consists of the following optimization problem:

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

Notation:

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

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 $\boldsymbol{\theta} \in \operatorname{supp} Q$,*

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

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

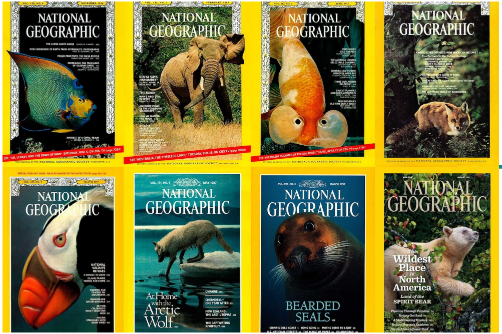
Relative Entropy Asymmetry

Definition (Generalized Relative Entropy)

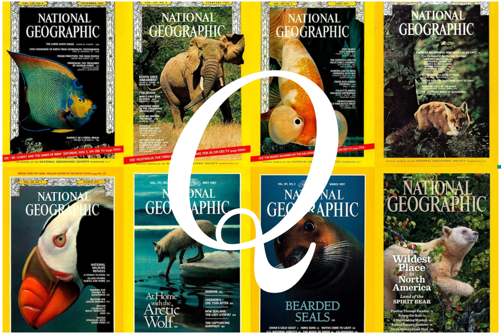
Given two σ-finite measures P and Q on the same measurable space, such that $P \ll Q$

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

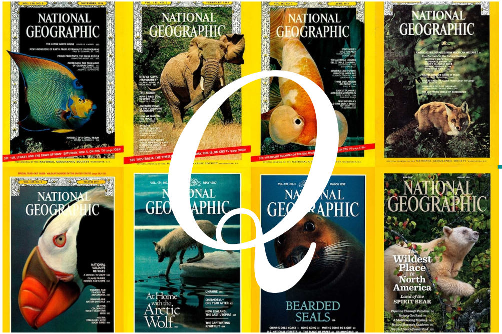
- Relative entropy is asymmetric: $\mathsf{D}(P\|Q) \neq \mathsf{D}(Q\|P)$
- For most cases of interest $P \ll Q \not\Rightarrow Q \ll P$
- Solution probability measure is **constrained** to $\mathrm{supp}(P) \subseteq \mathrm{supp}(Q)$



Prior Knowledge



Prior Knowledge

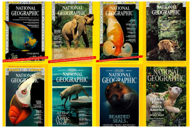


Prior Knowledge

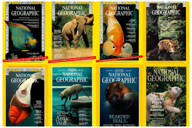
Set of All Models

$\mathcal{M}$

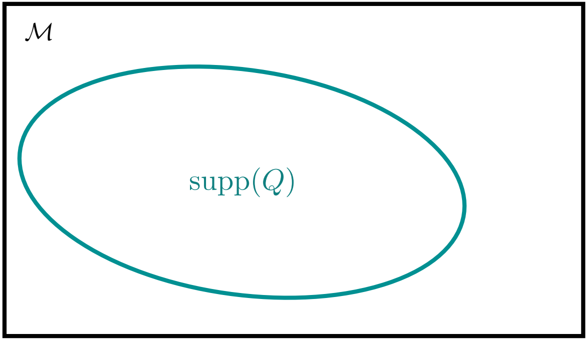
Prior Knowledge



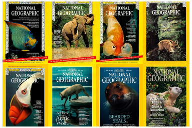
Prior Knowledge



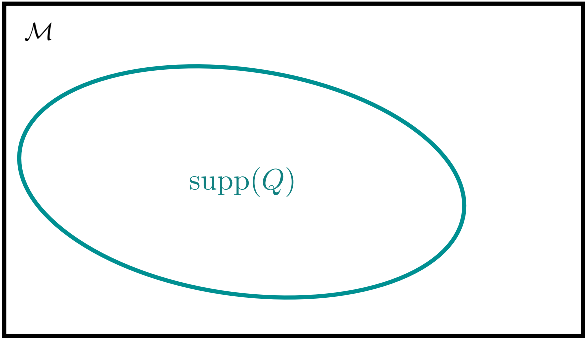
Set of All Models



Prior Knowledge

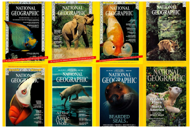


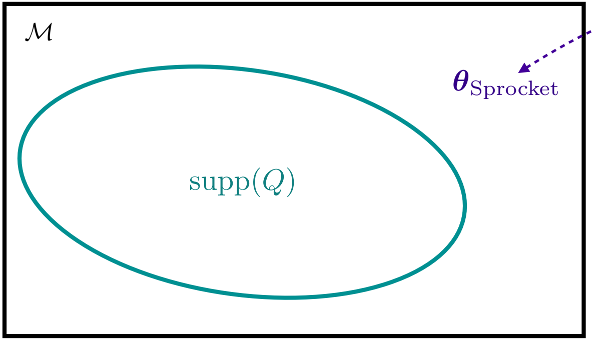
Set of All Models



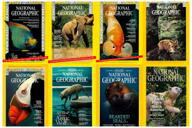
Set of All Models

Prior Knowledge

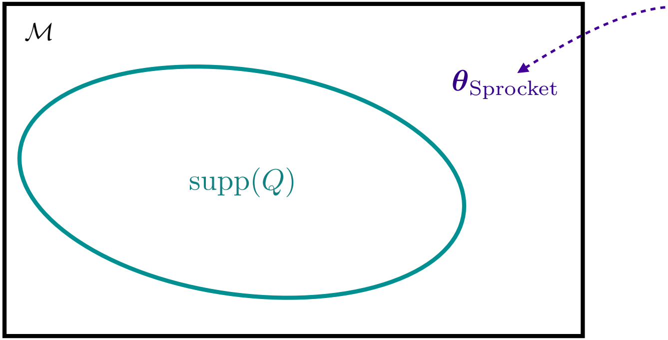




Prior Knowledge

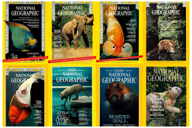


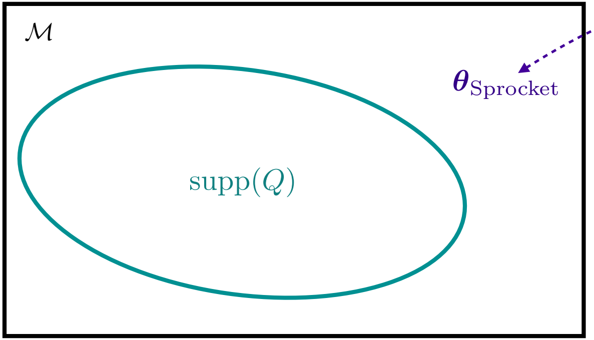
Set of All Models



Set of All Models

Prior Knowledge





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

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 $\triangledown_Q(\mathcal{M}, \mathscr{F}) \triangleq \{P \in \triangle(\mathcal{M}, \mathscr{F}) : Q \ll P\}$ given by

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

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

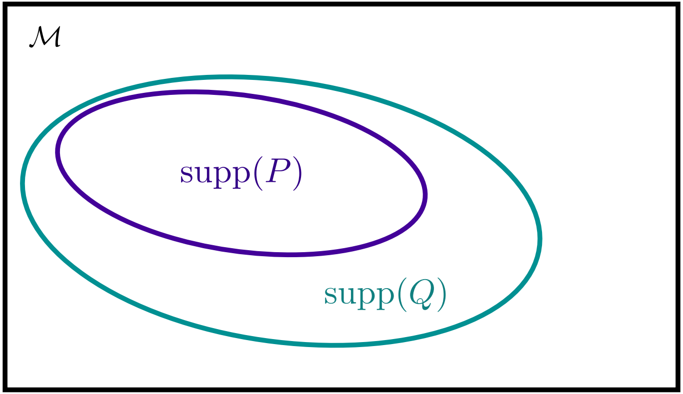
The ERM-RER Type-II problem, with parameters $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $\lambda \in (0, +\infty)$, consists of the optimization over the domain $\triangledown_Q(\mathcal{M}, \mathscr{F}) \triangleq \{P \in \triangle(\mathcal{M}, \mathscr{F}) : Q \ll P\}$ given by

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

- **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

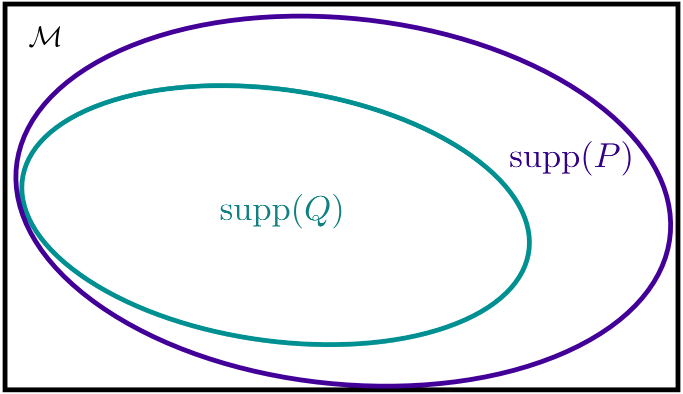
F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Analysis of the Relative Entropy Asymmetry in Regularized Empirical Risk Minimization," in *Proc. IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun. 2023.

Set of All Models



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

Set of All Models



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

Type-II ERM-RER Problem

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

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

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

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

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

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

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

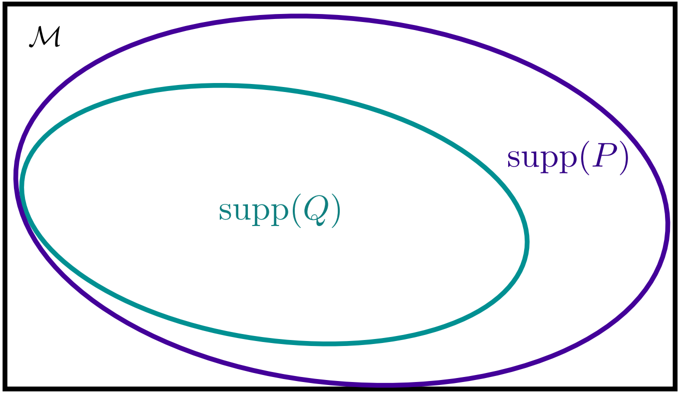
$$\int \frac{\lambda}{\beta + \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})} \mathrm{d}Q(\boldsymbol{\theta}) = 1,$$

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

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

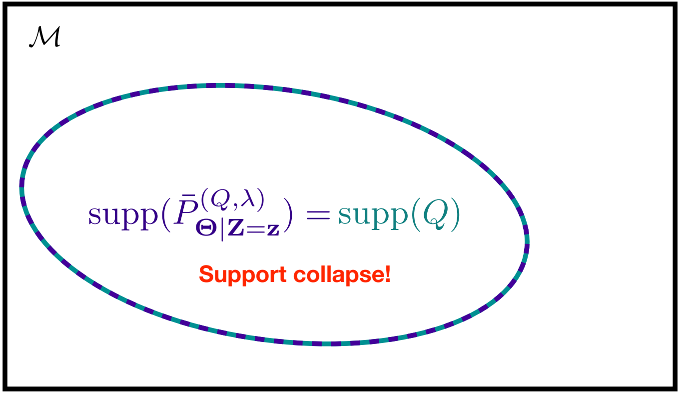
F. Daunas, I. Esnaola, S.M. Perlaza, and H.V. Poor, "Analysis of the Relative Entropy Asymmetry in Regularized Empirical Risk Minimization," in *Proc. IEEE International Symposium on Information Theory*, Taipei, Taiwan, Jun. 2023.

Set of All Models



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

Set of All Models



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

Type-II ERM-RER Problem

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**

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

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.

Type-II ERM-RER Problem

Brief Sketch of the Proof:

- Solve **ancillary problem**

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

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

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

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.

Type-II ERM-RER Problem

Brief Sketch of the Proof:

- Solve **ancillary problem**

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

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

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

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.

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

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

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

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

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

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

- 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 $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$.

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 $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$.

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

Link between expected empirical risk and normalization function:

Lemma

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

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

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

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

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

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:

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

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

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

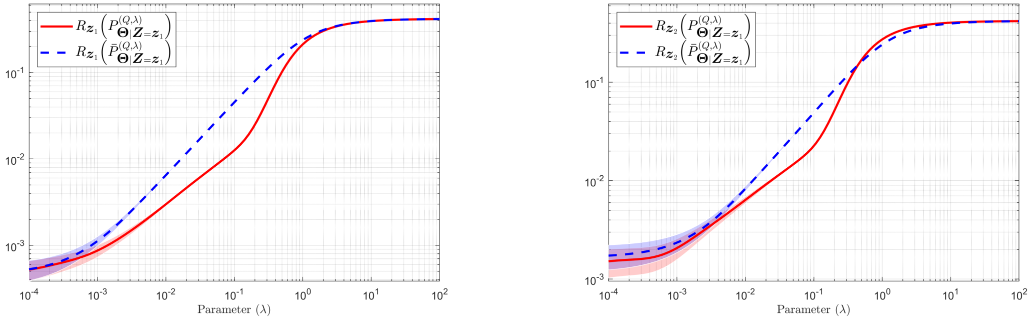
Type-I $\Rightarrow$ Type-II Equivalence:

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

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

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

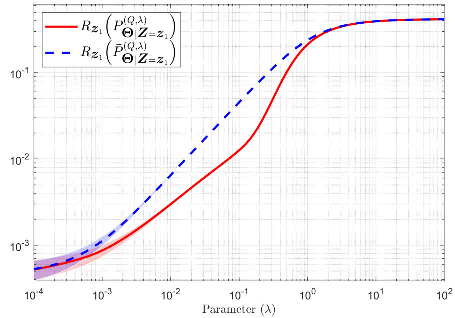
We train a **binary classifier** to distinguish '`six`' and '`seven`' in the MNIST dataset with the ERM-RER Type-I and Type-II



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



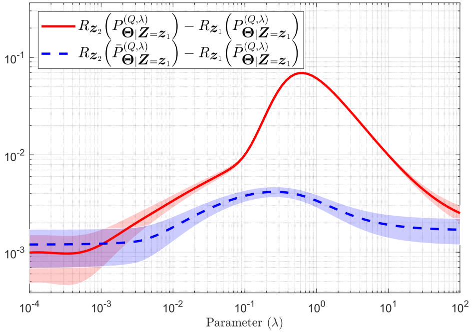


Table of Contents

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

ERM wih f-divergence Regularization

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

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

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

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

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)

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

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

- **Remark:** $\sqrt{\mathrm{JS}\,(P, Q)}$ 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

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

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

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

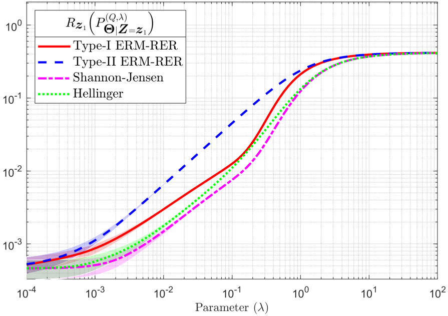
- The link to f-divergence characterization is

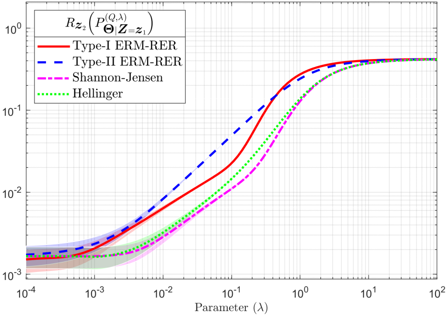
$$f(x) = (x-1)^2,$$
$$\dot{f}(x) = 2(x-1).$$

- The ERM-fDR solution yields

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

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

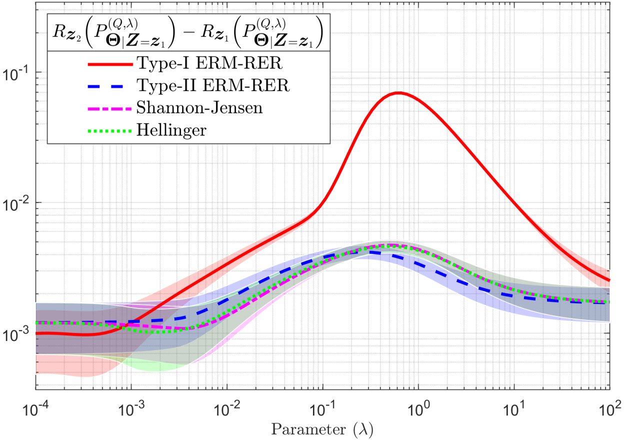




Numerical Comparison of Several Regularizations

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

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.