

Machine Learning Regularization via Information Measures

Iñaki Esnaola

`esnaola@sheffield.ac.uk`

School of Electrical and Electronic Engineering
University of Sheffield

Selected Lectures on Statistical Machine Learning

Inria Centre d'Université Côte d'Azur

February 25, 2026

Table of Contents

Regularization in Empirical Risk Minimization

- 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

Legislation for Trustworthy and Secure AI

<https://artificialintelligenceact.eu>



**EU Artificial
Intelligence Act**

[The Act](#) ▾

[Implementation](#) ▾

[Compliance](#) ▾

[About us](#)

[Newsletter](#) ↗

 [EN](#) ▾

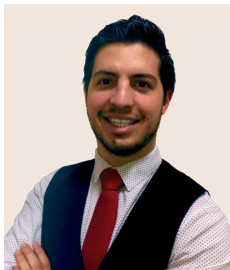
The EU Artificial Intelligence Act

Up-to-date developments and analyses of the EU AI Act

Legislation for Trustworthy and Secure AI

<https://artificialintelligenceact.eu>

- ▶ The EU has produced the first **common regulatory framework**: **Artificial Intelligence Act (AIA)**
- ▶ **Summary:**
 - ▶ **Prohibited AI systems:** biometric categorisation and identification, human emotion identification, manipulation of human emotion, social scoring.
 - ▶ **High-risk AI Systems:** operation in domains such as essential public infrastructure, water, gas and electricity supplies, education, law enforcement, and insurance banking.
 - ▶ **Obligations** for organisations operating on these domains, including:
 - ▶ **Transparency:** must provide instructions for use detailing characteristics, **capabilities, and limitations of the system.**
 - ▶ **Accuracy, robustness and cybersecurity:** resilient to errors or inconsistencies and resilient to cyber-attacks.
- ▶ **USA: Blueprint for an AI Bill of Rights** (October 2022) (Not binding)
<https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/>
- ▶ **UK: A pro-innovation approach to AI regulation** (March 2023)
<https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper>



Francisco Daunas
*University of Sheffield,
United Kingdom*



Samir Perlaza
*Centre INRIA d'Université
Côte d'Azur, France*



Gaetan Bisson
*University of French Polynesia,
French Polynesia*



H. Vincent Poor
*Princeton University,
United States*

Table of Contents

Regularization in Empirical Risk Minimization

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

Supervised Learning

Notation and Terminology

Consider the following **supervised learning** setting:

- ▶ **Elements:** set $\mathcal{X}$ of *patterns*, set $\mathcal{Y}$ of *labels*, and set $\mathcal{M} \subset \mathbb{R}^d$ of *models* with $d \in \mathbb{N}$.
- ▶ **Hypothesis:** function $f : \mathcal{M} \times \mathcal{X} \rightarrow \mathcal{Y}$.

Statistical Assumptions: *Realizability*

Two random variables X and Y satisfy

$$Y = f(\theta^*, X), \tag{1}$$

for some specific model θ^* (*optimal* model or hypothesis).

- ▶ Model θ^* is **unknown**: learner knowledge of the model is through *dataset*.
- ▶ *Dataset* $\mathbf{z} = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ **is available** to learner.

Supervised Learning

Notation and Terminology

Consider the following **supervised learning** setting:

- **Elements:** set $\mathcal{X}$ of *patterns*, set $\mathcal{Y}$ of *labels*, and set $\mathcal{M} \subset \mathbb{R}^d$ of *models* with $d \in \mathbb{N}$.
- **Hypothesis:** function $f : \mathcal{M} \times \mathcal{X} \rightarrow \mathcal{Y}$.

Statistical Assumptions: *Realizability*

Two random variables X and Y satisfy

$$Y = f(\theta^*, X), \tag{1}$$

for some specific model θ^* (*optimal* model or hypothesis).

Objective of the learner: Model Selection

Given a dataset $z \in (\mathcal{X} \times \mathcal{Y})^n$, find the model θ^* in (1).

Supervised Learning

Empirical Risk Minimization

Challenge: the learner receives a dataset sampled from an unknown probability distribution.

Definition (Risk and Loss)

Given the entry in the dataset $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the model $\theta \in \mathcal{M}$ induces the *risk* $\ell(f(\theta, x), y)$, where $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, +\infty)$ is a *loss* function.

Definition (Empirical Risk)

Given the dataset $\mathbf{z} = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, the *empirical risk* induced by the model $\theta \in \mathcal{M}$ is

$$\mathbf{L}_{\mathbf{z}}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(f(\theta, x_i), y_i).$$

Supervised Learning

Empirical Risk Minimization

Challenge: the learner receives a dataset sampled from an unknown probability distribution.

Definition (Risk and Loss)

Given the entry in the dataset $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the model $\theta \in \mathcal{M}$ induces the *risk* $\ell(f(\theta, x), y)$, where $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, +\infty)$ is a *loss* function.

Definition (Empirical Risk)

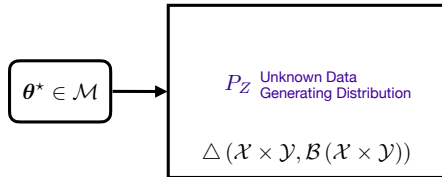
Given the dataset $z = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, the *empirical risk* induced by the model $\theta \in \mathcal{M}$ is

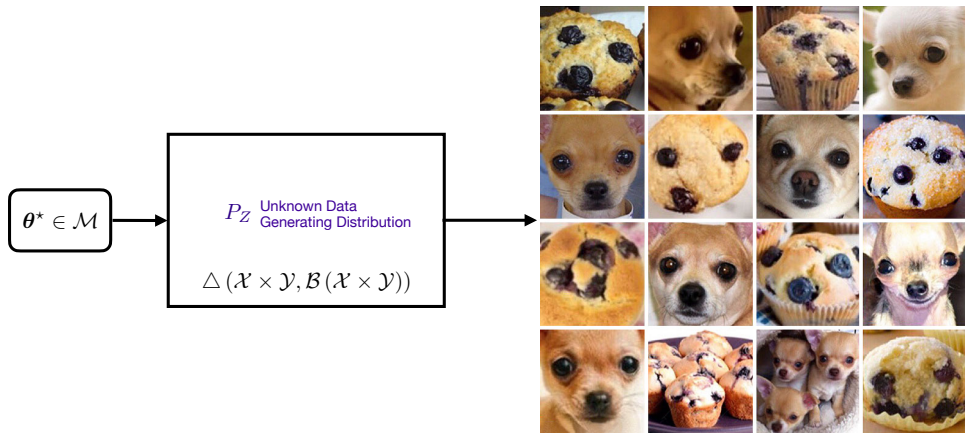
$$L_z(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(f(\theta, x_i), y_i).$$

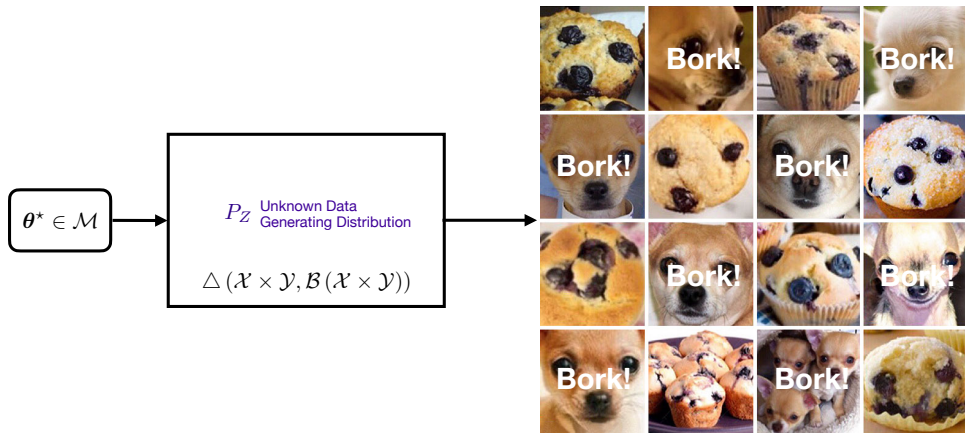
Problem Formulation: Empirical Risk Minimization (ERM)

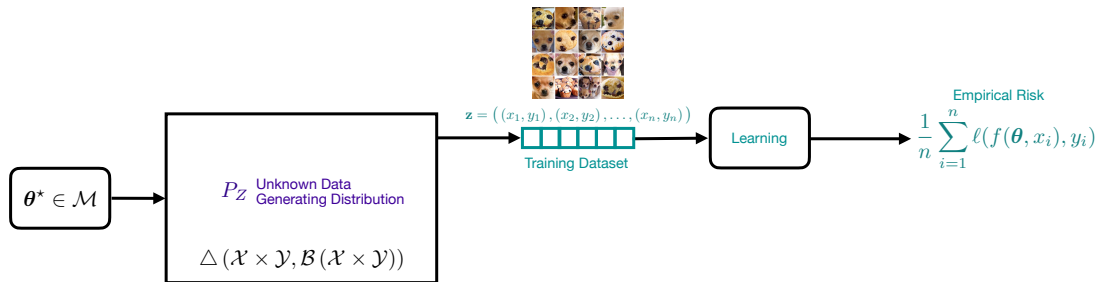
Given the dataset z , the ERM problem is

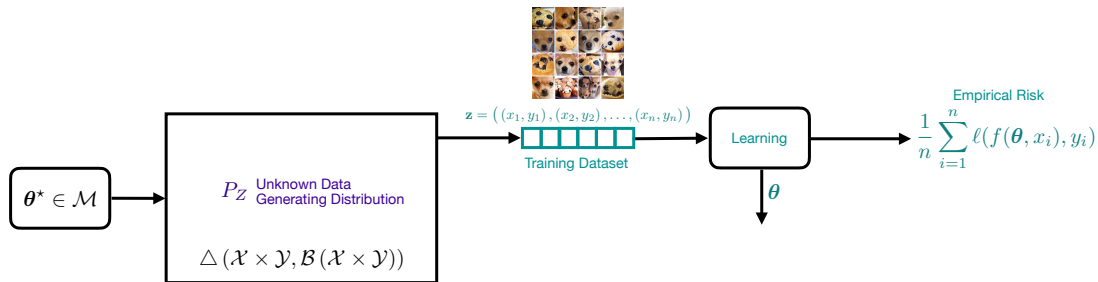
$$\min_{\theta \in \mathcal{M}} L_z(\theta).$$

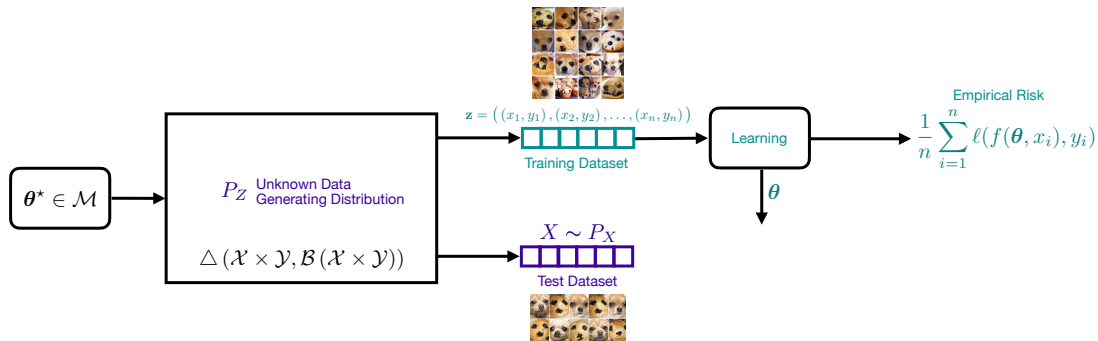


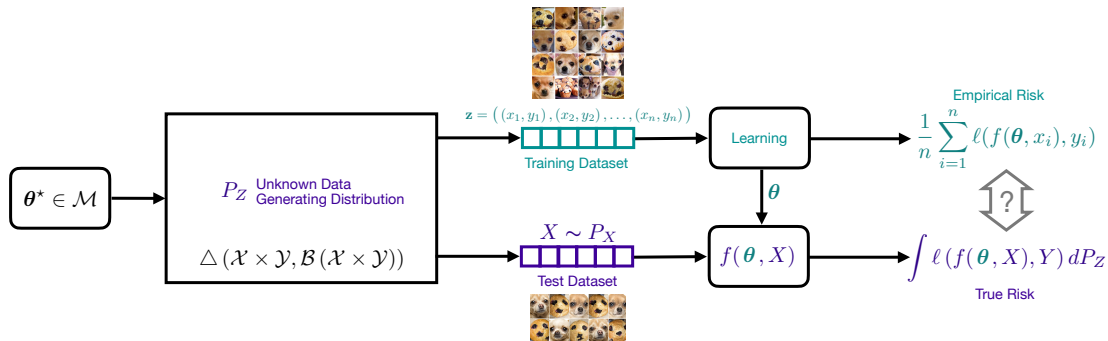












The Problem of Supervised Learning

Empirical Risk Minimization

Problem Formulation: Empirical Risk Minimization (ERM)

Given the dataset z , the ERM problem is

$$\min_{\theta \in \mathcal{M}} L_z(\theta).$$

Set of **solutions to the ERM problem**:

$$\mathcal{T}(z) \triangleq \arg \min_{\theta \in \mathcal{M}} L_z(\theta).$$

The Problem of Supervised Learning

Empirical Risk Minimization

Problem Formulation: Empirical Risk Minimization (ERM)

Given the dataset z , the ERM problem is

$$\min_{\theta \in \mathcal{M}} L_z(\theta).$$

Is this a *good* learning strategy for general loss functions?

- ▶ Hypothesis classes for which *uniform convergence* holds are **PAC learnable** [Dudley et al., 1991].
- ▶ For binary classification problems **PAC learnability** implies uniform convergence but this is not the case for more general problems [Shalev-Shwartz et al., 2010].
- ▶ **VC dimension** establishes the link between PAC learnability and uniform convergence [Vapnik and Chervonenkis, 1971, Blumer et al., 1989].
- ▶ **Rademacher complexity** based generalization guarantees [Shalev-Shwartz and Ben-David, 2014].

Many of these techniques rely on *worst-case* measures over the hypothesis class: **overly pessimistic**.

ERM with Relative Entropy Regularization (ERM-RER)

Definition (Expected Empirical Risk)

Given a dataset $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $R_{\mathbf{z}} : \Delta(\mathcal{M}, \mathcal{B}(\mathcal{M})) \rightarrow [0, +\infty)$ be such that for all σ -finite measures $P \in \Delta(\mathcal{M}, \mathcal{B}(\mathcal{M}))$, it holds that

$$R_{\mathbf{z}}(P) = \int L_{\mathbf{z}}(\boldsymbol{\theta}) dP(\boldsymbol{\theta}).$$

When P is a probability measure, the **expected empirical risk** induced by P is $R_{\mathbf{z}}(P)$.

ERM with Relative Entropy Regularization (ERM-RER)

Definition (Expected Empirical Risk)

Given a dataset $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $R_{\mathbf{z}} : \Delta(\mathcal{M}, \mathcal{B}(\mathcal{M})) \rightarrow [0, +\infty)$ be such that for all σ -finite measures $P \in \Delta(\mathcal{M}, \mathcal{B}(\mathcal{M}))$, it holds that

$$R_{\mathbf{z}}(P) = \int L_{\mathbf{z}}(\boldsymbol{\theta}) dP(\boldsymbol{\theta}).$$

When P is a probability measure, the **expected empirical risk** induced by P is $R_{\mathbf{z}}(P)$.

Definition (Generalized Relative Entropy)

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

$$D(P\|Q) \triangleq \int \frac{dP}{dQ}(\boldsymbol{\theta}) \log \left(\frac{dP}{dQ}(\boldsymbol{\theta}) \right) dQ(\boldsymbol{\theta}).$$

ERM with Relative Entropy Regularization (ERM-RER)

[Catoni, 2007, Zhang, 2006, Asadi et al., 2018, Jaakkola et al., 1999, Xu and Raginsky, 2017]

Definition (Expected Empirical Risk)

Given a dataset $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $R_{\mathbf{z}} : \Delta(\mathcal{M}, \mathcal{B}(\mathcal{M})) \rightarrow [0, +\infty)$ be such that for all σ -finite measures $P \in \Delta(\mathcal{M}, \mathcal{B}(\mathcal{M}))$, it holds that

$$R_{\mathbf{z}}(P) = \int L_{\mathbf{z}}(\boldsymbol{\theta}) dP(\boldsymbol{\theta}).$$

When P is a probability measure, the **expected empirical risk** induced by P is $R_{\mathbf{z}}(P)$.

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

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

$$\min_{P \in \Delta_Q(\mathcal{M}, \mathcal{B}(\mathcal{M}))} R_{\mathbf{z}}(P) + \lambda D(P \| Q).$$

ERM with Relative Entropy Regularization (ERM-RER)

S.M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical Risk Minimization With Relative Entropy Regularization: Optimality and Sensitivity Analysis," in *Proc. IEEE International Symposium on Information Theory*, Helsinki, Finland, Jul. 2022

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

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

$$\min_{P \in \Delta_Q(\mathcal{M}, \mathcal{B}(\mathcal{M}))} R_z(P) + \lambda D(P \| Q).$$

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 incorporate **inductive bias** in a transfer learning setting (*structural risk minimization*) [Bu et al. 2021].

ERM with Relative Entropy Regularization (ERM-RER)

[Catoni, 2007, Zhang, 2006, Asadi et al., 2018, Jaakkola et al., 1999, Xu and Raginsky, 2017]

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

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

$$\min_{P \in \Delta_Q(\mathcal{M}, \mathcal{B}(\mathcal{M}))} R_z(P) + \lambda D(P \| Q).$$

Let the (*log-partition*) function $K_{Q,z} : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ be such that for all $t \in \mathbb{R}$,

$$K_{Q,z}(t) = \log \left(\int \exp(t L_z(\theta)) dQ(\theta) \right).$$

Let also the set $\mathcal{K}_{Q,z} \subset \mathbb{R}$ be

$$\mathcal{K}_{Q,z} \triangleq \left\{ \lambda \in (0, +\infty) : K_{Q,z} \left(-\frac{1}{\lambda} \right) < +\infty \right\}.$$

ERM with Relative Entropy Regularization (ERM-RER)

S.M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical Risk Minimization With Relative Entropy Regularization: Optimality and Sensitivity Analysis," in *Proc. IEEE International Symposium on Information Theory*, Helsinki, Finland, Jul. 2022

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

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

$$\min_{P \in \Delta_Q(\mathcal{M}, \mathcal{B}(\mathcal{M}))} R_z(P) + \lambda D(P \| Q).$$

Theorem

For all $\lambda \in \mathcal{K}_{Q,z}$, the solution to the ERM-RER problem, denoted by $P_{\Theta|Z=z}^{(Q,\lambda)} \in \Delta_Q(\mathcal{M}, \mathcal{B}(\mathcal{M}))$, is a **unique** measure whose Radon-Nikodym derivative with respect to Q satisfies for all $\theta \in \text{supp}(Q)$,

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = \exp\left(-K_{Q,z}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}L_z(\theta)\right).$$

Table of Contents

Regularization in Empirical Risk Minimization

- 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

Relative Entropy Asymmetry

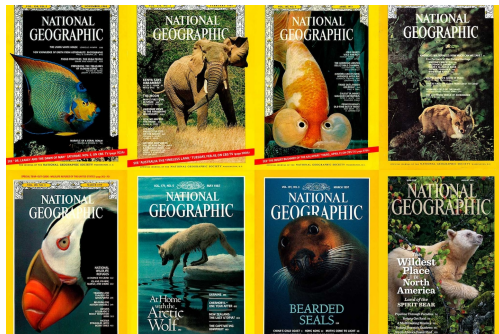
Definition (Generalized Relative Entropy)

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

$$D(P\|Q) \triangleq \int \frac{dP}{dQ}(\boldsymbol{\theta}) \log \left(\frac{dP}{dQ}(\boldsymbol{\theta}) \right) dQ(\boldsymbol{\theta}).$$

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





Prior Knowledge





Prior Knowledge





Prior Knowledge



Prior Knowledge



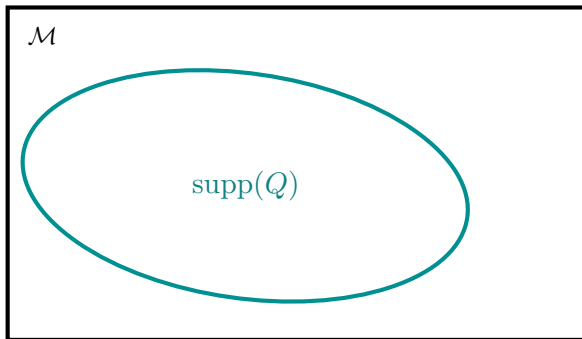
Set of All Models

$\mathcal{M}$

Prior Knowledge



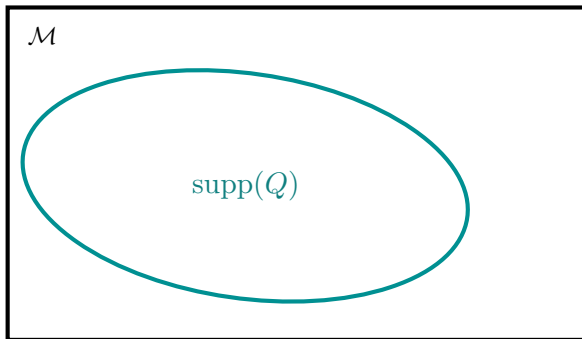
Set of All Models



Prior Knowledge



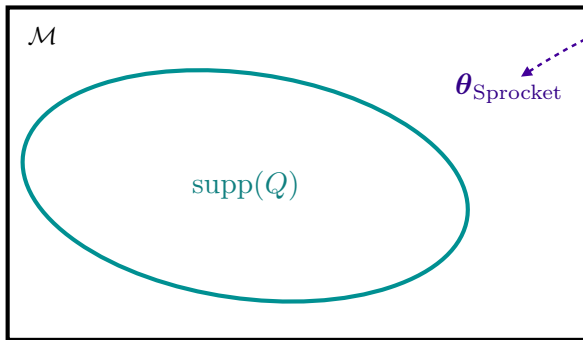
Set of All Models



Prior Knowledge



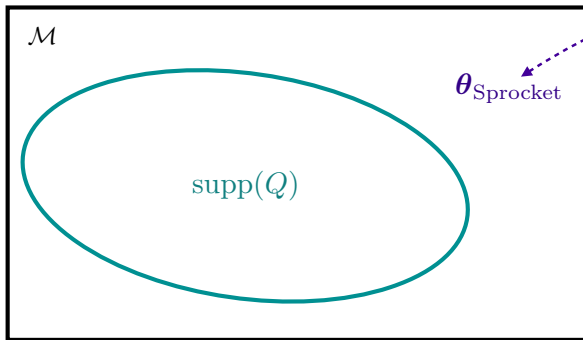
Set of All Models



Prior Knowledge



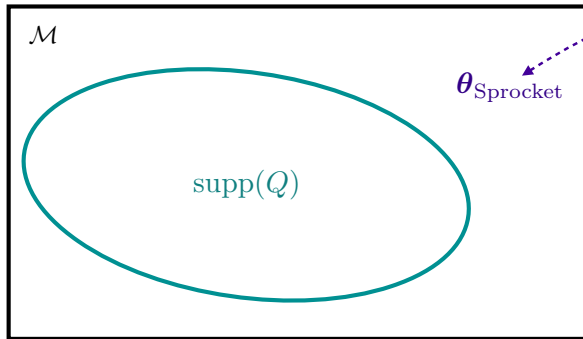
Set of All Models



Prior Knowledge



Set of All Models



$$\theta_{\text{Sprocket}} \notin \text{supp}(P)$$

Type-II ERM-RER Problem

Problem Formulation: Type-II ERM-RER

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

$$\min_{P \in \nabla_Q(\mathcal{M}, \mathcal{F})} R_{\mathbf{z}}(P) + \lambda D(Q \| P).$$

F. Daunas, I. Esnaola, S. M. Perlaza and H. Vincent Poor, "Asymmetry of the Relative Entropy in the Regularization of Empirical Risk Minimization," *IEEE Transactions on Information Theory*, vol. 71, no. 8, pp. 6198-6226, Aug. 2025

Type-II ERM-RER Problem

Theorem

If there exists a real β such that

$$\beta \in \{t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \text{supp} Q, 0 < t + L_z(\boldsymbol{\theta})\},$$

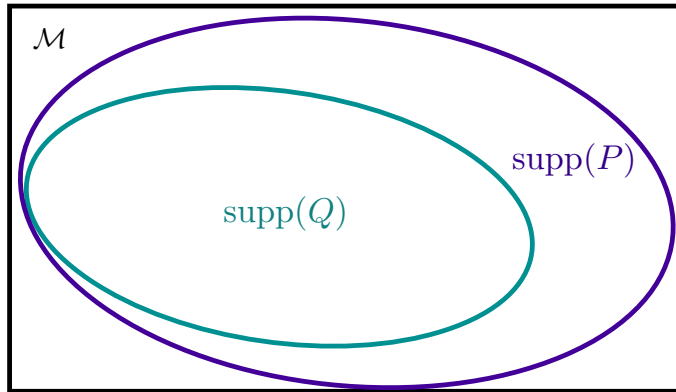
and

$$\int \frac{\lambda}{\beta + L_z(\boldsymbol{\theta})} dQ(\boldsymbol{\theta}) = 1,$$

then the unique solution to the **Type-II ERM-RER problem** is the probability measure $\bar{P}_{\boldsymbol{\Theta}|Z=z}^{(Q,\lambda)}$, which for all $\boldsymbol{\theta} \in \text{supp}(Q)$, the Radon-Nikodym derivative $\bar{P}_{\boldsymbol{\Theta}|Z=z}^{(Q,\lambda)}$ with respect to Q satisfies

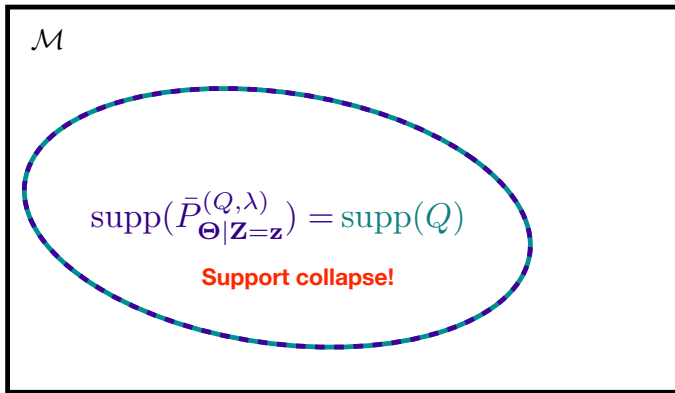
$$\frac{d\bar{P}_{\boldsymbol{\Theta}|Z=z}^{(Q,\lambda)}}{dQ}(\boldsymbol{\theta}) = \frac{\lambda}{\bar{K}_{Q,z}(\lambda) + L_z(\boldsymbol{\theta})}.$$

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. Vincent Poor, "Asymmetry of the Relative Entropy in the Regularization of Empirical Risk Minimization," *IEEE Transactions on Information Theory*, vol. 71, no. 8, pp. 6198-6226, Aug. 2025

Type-II ERM-RER Problem

Brief Sketch of the Proof:

- Solve **ancillary problem**

$$\min_{P \in \mathcal{O}_Q(\mathcal{M}, \mathcal{F})} R_z(P) + \lambda D(Q \| P), \quad \text{with} \quad \mathcal{O}_Q(\mathcal{M}, \mathcal{F}) \triangleq \nabla_Q(\mathcal{M}, \mathcal{F}) \cap \Delta_Q(\mathcal{M}, \mathcal{F})$$

Type-II ERM-RER Problem

Brief Sketch of the Proof:

- Solve **ancillary problem**

$$\min_{P \in \bigcirc_Q(\mathcal{M}, \mathcal{F})} R_z(P) + \lambda D(Q \| P), \quad \text{with} \quad \bigcirc_Q(\mathcal{M}, \mathcal{F}) \triangleq \nabla_Q(\mathcal{M}, \mathcal{F}) \cap \Delta_Q(\mathcal{M}, \mathcal{F})$$

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

$$\min_{V \in \nabla_Q(\mathcal{M}, \mathcal{F}) \setminus \bigcirc_Q(\mathcal{M}, \mathcal{F})} R_z(V) + \lambda D(Q \| V) > \min_{P \in \bigcirc_Q(\mathcal{M}, \mathcal{F})} R_z(P) + \lambda D(Q \| P).$$

Type-II ERM-RER Problem

Brief Sketch of the Proof:

- Solve **ancillary problem**

$$\min_{P \in \bigcirc_Q(\mathcal{M}, \mathcal{F})} R_z(P) + \lambda D(Q \| P), \quad \text{with} \quad \bigcirc_Q(\mathcal{M}, \mathcal{F}) \triangleq \nabla_Q(\mathcal{M}, \mathcal{F}) \cap \Delta_Q(\mathcal{M}, \mathcal{F})$$

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

$$\min_{V \in \nabla_Q(\mathcal{M}, \mathcal{F}) \setminus \bigcirc_Q(\mathcal{M}, \mathcal{F})} R_z(V) + \lambda D(Q \| V) > \min_{P \in \bigcirc_Q(\mathcal{M}, \mathcal{F})} R_z(P) + \lambda D(Q \| P).$$

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. Vincent Poor, "Asymmetry of the Relative Entropy in the Regularization of Empirical Risk Minimization," *IEEE Transactions on Information Theory*, vol. 71, no. 8, pp. 6198-6226, Aug. 2025

Type-II ERM-RER Properties

Normalization Function

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

$$\int \frac{\lambda}{\beta + L_z(\boldsymbol{\theta})} dQ(\boldsymbol{\theta}) = 1.$$

Type-II ERM-RER Properties

Normalization Function

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

$$\int \frac{\lambda}{\beta + L_z(\theta)} dQ(\theta) = 1.$$

Definition (Normalization Function)

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

Note that the Radon-Nikodym derivative of the solution is

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

Type-II ERM-RER Properties

Optimal models without regularization

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

$$\mathcal{L}_z(\delta) \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : L_z(\boldsymbol{\theta}) \leq \delta\}.$$

- Best achievable performance **without regularization**:

$$\delta_{Q,z}^* \triangleq \inf\{\delta \in [0, \infty) : Q(\mathcal{L}_z(\delta)) > 0\}.$$

- Solution models for the **Empirical Risk Minimization (withouth regularization)** problem:

$$\mathcal{L}_{Q,z}^* \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : L_z(\boldsymbol{\theta}) = \delta_{Q,z}^*\}.$$

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 $\theta \in \text{supp} Q$ it holds that

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

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

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 $\theta \in \text{supp} Q$ it holds that

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

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

Empirical risk dominates inductive bias for any regularization regime.

Lemma

For all $(\theta_1, \theta_2) \in (\text{supp} Q)^2$, such that $L_z(\theta_1) \leq L_z(\theta_2)$, it holds that

$$\frac{d\bar{P}_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta_2) \leq \frac{d\bar{P}_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta_1),$$

with equality if and only if $L_z(\theta_1) = L_z(\theta_2)$.

Type-II ERM-RER Properties

Asymptotes of the Radon-Nikodym Derivative

Type-II ERM-RER Properties

Asymptotes of the Radon-Nikodym Derivative

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

Lemma

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

Type-II ERM-RER Properties

Asymptotes of the Radon-Nikodym Derivative

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

Lemma

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

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

Lemma

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

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

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

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

Type-II ERM-RER Properties

Expected Empirical Risk

Link between expected empirical risk and normalization function:

Lemma

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

Lower bound on evidence provided by data compared to prior knowledge:

Lemma

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

Type-II ERM-RER Properties

Expected Empirical Risk

Link between expected empirical risk and normalization function:

Lemma

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

Lower bound on evidence provided by data compared to prior knowledge:

Lemma

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

Equivalence of Type-I and Type-II Regularization

Equivalence of Type-I and Type-II Regularization

Theorem

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

$$\min_{P \in \nabla_Q(\mathcal{M})} \int \mathbf{L}_{\mathbf{z}}(\boldsymbol{\theta}) dP(\boldsymbol{\theta}) + \lambda D(Q \| P) = \min_{P \in \Delta_Q(\mathcal{M})} \int \mathbf{V}_{Q, \mathbf{z}, \lambda}(\boldsymbol{\theta}) dP(\boldsymbol{\theta}) + D(P \| Q),$$

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

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

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

$$\min_{P \in \Delta_Q(\mathcal{M})} \int \mathbf{L}_{\mathbf{z}}(\boldsymbol{\theta}) dP(\boldsymbol{\theta}) + \lambda D(P \| Q) = \min_{P \in \nabla_Q(\mathcal{M})} \int \mathbf{W}_{Q, \mathbf{z}, \lambda}(\boldsymbol{\theta}) dP(\boldsymbol{\theta}) + D(Q \| P),$$

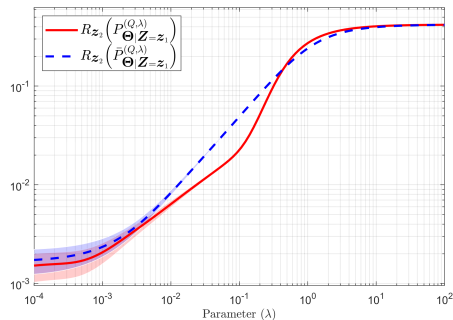
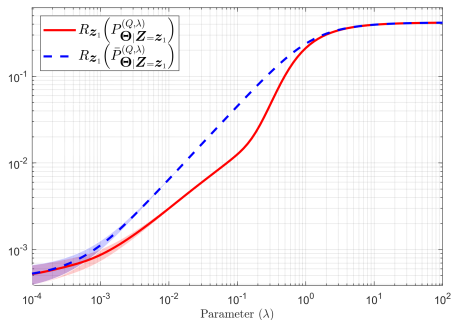
where the function $\mathbf{W}_{Q, \mathbf{z}, \lambda} : \mathcal{M} \rightarrow \mathbb{R}$ is defined as

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

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 dataset¹ with the ERM-RER Type-I and Type-II

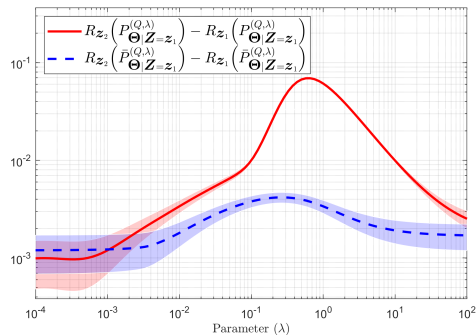
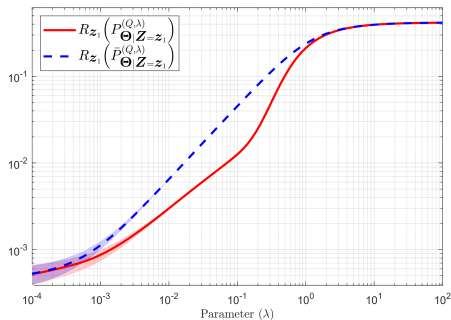


¹Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Aug. 1998.

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 dataset² with the ERM-RER Type-I and Type-II



²Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Aug. 1998.

Table of Contents

Regularization in Empirical Risk Minimization

- 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

f -divergences

Definition

Definition (f -divergence [Csiszár, 1967])

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

$$D_f(P\|Q) \triangleq \int f\left(\frac{dP}{dQ}(\boldsymbol{\theta})\right) dQ(\boldsymbol{\theta}),$$

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

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: $f(x) = \frac{1-x}{2x+2}$

Basic Properties

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

ERM with f -divergence Regularization

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

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

$$\min_{P \in \Delta_Q(\mathcal{M}, \mathcal{F})} R_{\mathbf{z}}(P) + \lambda D_f(P \| Q),$$

with optimization domain

$$\Delta_Q(\mathcal{M}, \mathcal{F}) \triangleq \{P \in \Delta(\mathcal{M}, \mathcal{F}) : P \ll Q\}.$$

P. Alquier, "Non-exponentially weighted aggregation: regret bounds for unbounded loss functions," in *Proc. International Conference on Machine Learning*, vol. 139, Jul. 2021, pp. 207–218.

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.

Solution to the ERM- f DR

The solutions to the ERM- f DR problems is presented under the following **assumptions**:

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

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

and

$$\int \dot{f}^{-1} \left(-\frac{\beta + L_{\mathbf{z}}(\boldsymbol{\theta})}{\lambda} \right) dQ(\boldsymbol{\theta}) = 1$$

- ▶ The function $L_{\mathbf{z}}$ is **separable** with respect to the probability measure Q

Solution to the ERM- f DR

The solutions to the ERM- f DR problems is presented under the following **assumptions**:

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

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

and

$$\int \dot{f}^{-1} \left(-\frac{\beta + L_{\mathbf{z}}(\boldsymbol{\theta})}{\lambda} \right) dQ(\boldsymbol{\theta}) = 1$$

- ▶ The function $L_{\mathbf{z}}$ is **separable** with respect to the probability measure Q

Definition (Separable Empirical Risk Function)

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

$$L_{\mathbf{z}}(\boldsymbol{\theta}_1) < c < L_{\mathbf{z}}(\boldsymbol{\theta}_2) < \infty.$$

Solution to the ERM- f DR

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.

Solution to the ERM- f DR

Theorem

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

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = f^{-1}\left(-\frac{\beta + L_z(\theta)}{\lambda}\right).$$

Remarks:

- ▶ Probability measures Q and $P_{\Theta|Z=z}^{(Q,\lambda)}$ 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 with f -divergence Regularization

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

ERM with f -divergence Regularization

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

Setting

$$\begin{aligned}f(x) &= x \log x, \\ \dot{f}(x) &= \log x + 1,\end{aligned}$$

results in

$$D_f(P\|Q) = \int f\left(\frac{dP}{dQ}(\boldsymbol{\theta})\right) dQ(\boldsymbol{\theta}) = \int \log\left(\frac{dP}{dQ}(\boldsymbol{\theta})\right) dP(\boldsymbol{\theta}).$$

The ERM- f DR solution yields

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\boldsymbol{\theta}) = \exp\left(-\frac{\beta + \mathbf{L}_z(\boldsymbol{\theta}) + \lambda}{\lambda}\right).$$

ERM with f -divergence Regularization

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

ERM with f -divergence Regularization

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

Setting

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

results in

$$D_f(P\|Q) = \int f\left(\frac{dP}{dQ}(\boldsymbol{\theta})\right) dQ(\boldsymbol{\theta}) = - \int \log\left(\frac{dP}{dQ}(\boldsymbol{\theta})\right) dQ(\boldsymbol{\theta}) = \int \log\left(\frac{dQ}{dP}(\boldsymbol{\theta})\right) dQ(\boldsymbol{\theta}).$$

The ERM- f DR solution yields

$$\frac{dP_{\boldsymbol{\Theta}|Z=z}^{(Q,\lambda)}}{dQ}(\boldsymbol{\theta}) = \frac{\lambda}{\beta + L_z(\boldsymbol{\theta})}.$$

ERM with f -divergence Regularization

Common Cases: Jensen-Shannon Divergence

ERM with f -divergence Regularization

Common Cases: Jensen-Shannon Divergence

Definition (Jensen-Shannon Divergence)

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

$$\text{JS}(P, Q) = D(P \| \frac{1}{2}(P + Q)) + D(Q \| \frac{1}{2}(P + Q)).$$

- **Remark:** $\sqrt{\text{JS}(P, Q)}$ 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- f DR solution yields

$$\frac{dP_{\Theta|Z=z}^{(Q, \lambda)}}{dQ}(\theta) = \frac{1}{2 \exp(\frac{\beta + L_z(\theta)}{\lambda}) - 1}.$$

ERM with f -divergence Regularization

Common Cases: χ^2 -divergence

ERM with f -divergence Regularization

Common Cases: χ^2 -divergence

Definition (χ^2 -divergence)

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

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

- The link to f -divergence characterization is

$$f(x) = (x - 1)^2,$$

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

- The ERM- f DR solution yields

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\boldsymbol{\theta}) = -\frac{\beta + L_z(\boldsymbol{\theta})}{\lambda}.$$

ERM with f -divergence Regularization

Common Cases: Total Variation

ERM with f -divergence Regularization

Common Cases: Total Variation

Definition (Total Variation)

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

$$\text{TV}(P\|Q) = \frac{1}{2} \int \left| \frac{dP}{dQ}(\boldsymbol{\theta}) - 1 \right| dQ(\boldsymbol{\theta}).$$

- **Remark:** $\text{TV}(P\|Q)$ is a metric in the space of probability measure.
- The link to f -divergence characterization is

$$f(x) = \frac{1}{2}|x - 1|,$$
$$\dot{f}(x) = \frac{1}{2} \text{sign}(x - 1).$$

- The ERM- f DR solution yields

$$\frac{1}{2} \text{sign} \left(\frac{dP_{\boldsymbol{\Theta}|Z=z}^{(Q,\lambda)}}{dQ}(\boldsymbol{\theta}) \right) = -\frac{\beta + L_z(\boldsymbol{\theta})}{\lambda}.$$

ERM with f -divergence Regularization

Toy Example: Total Variation

ERM with f -divergence Regularization

Toy Example: Total Variation

Setting

- ▶ Let $\mathcal{M} = \{\theta_1, \theta_2\}$
- ▶ Let $\sum_{i=1}^2 Q(\theta_i) = 1$
- ▶ For $i \in \{1, 2\}$ it holds that $L_z(\theta_i) \geq 0$
- ▶ For $P \in \Delta(\mathcal{M}, \mathcal{F})$, it holds that $R_z(P) = L_z(\theta_1)P(\theta_1) + L_z(\theta_2)(1 - P(\theta_1))$
- ▶ Regularization term: $TV(P, Q) = |P(\theta_1) - Q(\theta_1)|$
- ▶ Resulting optimization problem:

$$\begin{aligned} & \min_{P \in \Delta_Q(\mathcal{M}, \mathcal{F})} R_z(P) + \lambda D_f(P \| Q) \\ &= \min_{P(\theta_1) \in [0, 1]} (L_z(\theta_1) - L_z(\theta_2))P(\theta_1) + L_z(\theta_2) + \lambda |P(\theta_1) - Q(\theta_1)|. \end{aligned}$$

ERM with f -divergence Regularization

Toy Example: Total Variation

ERM with f -divergence Regularization

Toy Example: Total Variation

Cost function is broken into two cases

$$\begin{aligned} & (\mathcal{L}_z(\theta_1) - \mathcal{L}_z(\theta_2))P(\theta_1) + \mathcal{L}_z(\theta_2) + \lambda |P(\theta_1) - Q(\theta_1)| \\ &= \begin{cases} (\mathcal{L}_z(\theta_1) - \mathcal{L}_z(\theta_2) + \lambda)P(\theta_1) + \mathcal{L}_z(\theta_2) - \lambda Q(\theta_1) & \text{for } P(\theta_1) \geq Q(\theta_1) \\ (\mathcal{L}_z(\theta_1) - \mathcal{L}_z(\theta_2) - \lambda)P(\theta_1) + \mathcal{L}_z(\theta_2) + \lambda Q(\theta_1) & \text{for } P(\theta_1) < Q(\theta_1). \end{cases} \end{aligned}$$

Remarks on the solution to the minimization problem

- ▶ For $\lambda = \mathcal{L}_z(\theta_1) - \mathcal{L}_z(\theta_2)$ and $\lambda = \mathcal{L}_z(\theta_2) - \mathcal{L}_z(\theta_1)$ *strict convexity* does not hold and there exist infinitely many solutions.
- ▶ For other values of λ two distinct cases emerge:
 - ▶ Solution assigns all mass to the boundary distribution, i.e. $P(\theta_1) = 0$ or $P(\theta_1) = 1$
 - ▶ Solution is $P = Q$

Revisiting the Regularization Equivalence

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?**

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 the definition of f -divergences. If the ERM- f DR problem possesses a solution, then

$$\min_{P \in \Delta_Q(\mathcal{M})} \int \mathbf{L}_z(\boldsymbol{\theta}) dP(\boldsymbol{\theta}) + \lambda D_f(P \| Q) = \min_{P \in \Delta_Q(\mathcal{M})} \int v(\mathbf{L}_z(\boldsymbol{\theta})) dP(\boldsymbol{\theta}) + \lambda D_g(P \| Q),$$

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

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

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

Example: Recovering the Equivalence Between Type-I and Type-II Regularization

- **Type-I:** $g(t) = t \log(t)$

$$\min_{P \in \Delta_Q(\mathcal{M})} \int v(\mathbf{L}_z(\boldsymbol{\theta})) dP(\boldsymbol{\theta}) + \lambda \underbrace{\mathbf{D}_g(P \| Q)}_{= \mathbf{D}(P \| Q)}$$

- **Type-II:** $f(t) = -\log(t)$

$$\min_{P \in \Delta_Q(\mathcal{M})} \int v(\mathbf{L}_z(\boldsymbol{\theta})) dP(\boldsymbol{\theta}) + \lambda \underbrace{\mathbf{D}_f(P \| Q)}_{= \mathbf{D}(Q \| P)}$$

- Regularization **invariance** under expected empirical risk transformation

Example: Recovering the Equivalence Between Type-I and Type-II Regularization

- **Type-I:** $g(t) = t \log(t)$

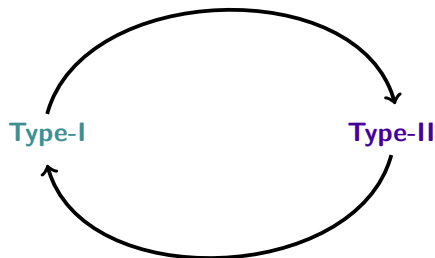
$$\min_{P \in \Delta_Q(\mathcal{M})} \int v(\mathbf{L}_z(\boldsymbol{\theta})) dP(\boldsymbol{\theta}) + \underbrace{\lambda \mathbf{D}_g(P \| Q)}_{=D(P \| Q)}$$

- **Type-II:** $f(t) = -\log(t)$

$$\min_{P \in \Delta_Q(\mathcal{M})} \int v(\mathbf{L}_z(\boldsymbol{\theta})) dP(\boldsymbol{\theta}) + \underbrace{\lambda \mathbf{D}_f(P \| Q)}_{=D(Q \| P)}$$

- Regularization **invariance** under expected empirical risk transformation

$$v(\mathbf{L}_z(\boldsymbol{\theta})) = \lambda \log(N'_{Q,z}(\lambda) + \mathbf{L}_z(\boldsymbol{\theta}))$$



$$\bar{v}(\mathbf{L}_z(\boldsymbol{\theta})) = \frac{\lambda}{\exp(-\frac{\mathbf{L}_z(\boldsymbol{\theta})}{\lambda}) - \log(\int \exp(-\frac{\mathbf{L}_z(\nu)}{\lambda}) dQ(\nu))} - N'_{Q,z}(\lambda)$$

Conclusions

Conclusions

- ▶ **Information type-measures** open the door to novel **operationally meaningful** regularization approaches.

Conclusions

- ▶ **Information type-measures** open the door to novel **operationally meaningful** regularization approaches.
- ▶ New tools for **generalization** analysis that open the door to exact characterizations.

Conclusions

- ▶ **Information type-measures** open the door to novel **operationally meaningful** regularization approaches.
- ▶ New tools for **generalization** analysis that open the door to exact characterizations.
- ▶ Equivalence results for f -divergence regularization unveil **link** between **loss** definition and **inductive bias** introduced by regularization.

Conclusions

- ▶ **Information type-measures** open the door to novel **operationally meaningful** regularization approaches.
- ▶ New tools for **generalization** analysis that open the door to exact characterizations.
- ▶ Equivalence results for f -divergence regularization unveil **link** between **loss** definition and **inductive bias** introduced by regularization.
- ▶ **Open problem:** overcoming the **inductive bias** introduced by these regularizers when data provides new evidence.

Conclusions

- ▶ **Information type-measures** open the door to novel **operationally meaningful** regularization approaches.
- ▶ New tools for **generalization** analysis that open the door to exact characterizations.
- ▶ Equivalence results for f -divergence regularization unveil **link** between **loss** definition and **inductive bias** introduced by regularization.
- ▶ **Open problem:** overcoming the **inductive bias** introduced by these regularizers when data provides new evidence.
- ▶ **Adapting** the f -divergence to different learning frameworks suggests **tailored regularizer designs**
 - ▶ Loss function definition
 - ▶ Model set adaptation to practical implementations
 - ▶ Robustness to dataset features

Conclusions

- ▶ **Information type-measures** open the door to novel **operationally meaningful** regularization approaches.
- ▶ New tools for **generalization** analysis that open the door to exact characterizations.
- ▶ Equivalence results for f -divergence regularization unveil **link** between **loss** definition and **inductive bias** introduced by regularization.
- ▶ **Open problem:** overcoming the **inductive bias** introduced by these regularizers when data provides new evidence.
- ▶ **Adapting** the f -divergence to different learning frameworks suggests **tailored regularizer designs**
 - ▶ Loss function definition
 - ▶ Model set adaptation to practical implementations
 - ▶ Robustness to dataset features
- ▶ **Open problem:** f -divergences that induce loss of **strict convexity** require novel approaches

Thanks!

`esnaola@sheffield.ac.uk`

Bibliography I



Asadi, A., Abbe, E., and Verdú, S. (2018).

Chaining mutual information and tightening generalization bounds.

In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, volume 31, Montréal, Canada.



Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989).

Learnability and the vapnik-chervonenkis dimension.

Journal of the ACM (JACM), 36(4):929–965.



Catoni, O. (2007).

PAC-Bayesian supervised classification: The thermodynamics of statistical learning, volume 56.

Institute of Mathematical Statistics Lecture Notes - Monograph Series, Beachwood, OH, USA, first edition.

Bibliography II

-  Csiszár, I. (1967).
Information-type measures of difference of probability distributions and indirect observation.
Studia Scientiarum Mathematicarum Hungarica, 2(1):299–318.
-  Dudley, R. M., Giné, E., and Zinn, J. (1991).
Uniform and universal Glivenko-Cantelli classes.
Journal of Theoretical Probability, 4(3):485–510.
-  Jaakkola, T., Meila, M., and Jebara, T. (1999).
Maximum entropy discrimination.
In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada.

Bibliography III



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

Understanding machine learning: From theory to algorithms.

Cambridge University Press, New York, NY, USA, first edition.



Shalev-Shwartz, S., Srebro, N., and Zhang, T. (2010).

Trading accuracy for sparsity in optimization problems with sparsity constraints.

SIAM Journal on Optimization, 20(6):2807–2832.



Vapnik, V. N. and Chervonenkis, A. Y. (1971).

On the uniform convergence of relative frequencies of events to their probabilities.

Theory of Probability and its applications, 16(2):264–279.

Bibliography IV

-  Xu, A. and Raginsky, M. (2017).
Information-theoretic analysis of generalization capability of learning algorithms.
In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, Long Beach, CA, USA.
-  Zhang, T. (2006).
Information-theoretic upper and lower bounds for statistical estimation.
IEEE Transactions on Information Theory, 52(4):1307–1321.

Bibliography I



Asadi, A., Abbe, E., and Verdú, S. (2018).

Chaining mutual information and tightening generalization bounds.

In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, volume 31, Montréal, Canada.



Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989).

Learnability and the vapnik-chervonenkis dimension.

Journal of the ACM (JACM), 36(4):929–965.



Catoni, O. (2007).

PAC-Bayesian supervised classification: The thermodynamics of statistical learning, volume 56.

Institute of Mathematical Statistics Lecture Notes - Monograph Series, Beachwood, OH, USA, first edition.

Bibliography II

-  Csiszár, I. (1967).
Information-type measures of difference of probability distributions and indirect observation.
Studia Scientiarum Mathematicarum Hungarica, 2(1):299–318.
-  Dudley, R. M., Giné, E., and Zinn, J. (1991).
Uniform and universal Glivenko-Cantelli classes.
Journal of Theoretical Probability, 4(3):485–510.
-  Jaakkola, T., Meila, M., and Jebara, T. (1999).
Maximum entropy discrimination.
In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada.

Bibliography III



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

Understanding machine learning: From theory to algorithms.

Cambridge University Press, New York, NY, USA, first edition.



Shalev-Shwartz, S., Srebro, N., and Zhang, T. (2010).

Trading accuracy for sparsity in optimization problems with sparsity constraints.

SIAM Journal on Optimization, 20(6):2807–2832.



Vapnik, V. N. and Chervonenkis, A. Y. (1971).

On the uniform convergence of relative frequencies of events to their probabilities.

Theory of Probability and its applications, 16(2):264–279.

Bibliography IV

-  Xu, A. and Raginsky, M. (2017).
Information-theoretic analysis of generalization capability of learning algorithms.
In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, Long Beach, CA, USA.
-  Zhang, T. (2006).
Information-theoretic upper and lower bounds for statistical estimation.
IEEE Transactions on Information Theory, 52(4):1307–1321.

Type-II ERM-RER Problem

Brief Sketch of the Proof:

- Solve **ancillary problem**

$$\min_{P \in \bigcirc_Q(\mathcal{M}, \mathcal{F})} R_z(P) + \lambda D(Q \| P), \quad \text{with} \quad \bigcirc_Q(\mathcal{M}, \mathcal{F}) \triangleq \nabla_Q(\mathcal{M}, \mathcal{F}) \cap \Delta_Q(\mathcal{M}, \mathcal{F})$$

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

$$\min_{V \in \nabla_Q(\mathcal{M}, \mathcal{F}) \setminus \bigcirc_Q(\mathcal{M}, \mathcal{F})} R_z(V) + \lambda D(Q \| V) > \min_{P \in \bigcirc_Q(\mathcal{M}, \mathcal{F})} R_z(P) + \lambda D(Q \| P).$$

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 $\mathcal{A}_{Q,z} \subseteq (0, \infty)$ and $\mathcal{C}_{Q,z} \subset \mathbb{R}$ be such that if $\lambda \in \mathcal{A}_{Q,z}$, then there exists a $\beta \in \mathcal{C}_{Q,z}$ that satisfies $\beta \in \{t \in \mathbb{R} : \forall \theta \in \text{supp} Q, 0 < t + L(z, \theta)\}$ and

$$\int \frac{\lambda}{\beta + L(z, \theta)} dQ(\theta) = 1.$$

Definition (Normalization Function)

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

Note that the Radon-Nikodym derivative of the solution is

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

Type-II ERM-RER Properties

Optimal models without regularization

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

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

- Best achievable performance **without regularization**:

$$\delta_{Q,z}^* \triangleq \inf\{\delta \in [0, \infty) : Q(\mathcal{L}_z(\delta)) > 0\}.$$

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

$$\mathcal{L}_{Q,z}^* \triangleq \{\theta \in \mathcal{M} : L(z, \theta) = \delta_{Q,z}^*\}.$$

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 $\theta \in \text{supp}Q$ it holds that

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

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

Empirical risk dominates inductive bias for any regularization regime.

Lemma

For all $(\theta_1, \theta_2) \in (\text{supp}Q)^2$, such that $L(z, \theta_1) \leq L(z, \theta_2)$, it holds that

$$\frac{d\bar{P}_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta_2) \leq \frac{d\bar{P}_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta_1),$$

with equality if and only if $L(z, \theta_1) = L(z, \theta_2)$.

Type-II ERM-RER Properties

Asymptotes of the Radon-Nikodym Derivative

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

Lemma

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

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

Lemma

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

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