The Role of Relative Entropy in Supervised Machine Learning

Samir M. Perlaza
Inria, Sophia Antipolis, France.

Electrical and Computer Engineering Department
Princeton University
Princeton, NJ, July 28, 2022

Research reported in:

Samir M. Perlaza, Iñaki Esnaola, and H. Vincent Poor. "**Sensitivity of the Gibbs Algorithm to Data Aggregation in Supervised Machine Learning**". Research Report, INRIA, No. RR-9474, Sophia Antipolis, France, Jun., 2022.

Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, Stefano Rini, "**Empirical Risk Minimization with Relative Entropy Regularizations**". Research Report, INRIA, No. RR-9454, Sophia Antipolis, France, Feb., 2022.

Table of Contents

Table of Contents

The Problem of Supervised Learning

Consider the following **supervised learning** scenario:

- three sets $\mathcal{X}$ (patterns), $\mathcal{Y}$ (labels) and $\mathcal{M} \subset \mathbb{R}^d$ (models), with $d \in \mathbb{N}$.
- a function $f : \mathcal{M} \times \mathcal{X} \to \mathcal{Y}$ (explicit expression **is known**)

Statistical Assumptions

Two random variables X and Y satisfy

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

for some specific model $\boldsymbol{\theta}^\star$ (optimal model or hypothesis).

- model $\boldsymbol{\theta}^\star$ is **unknown**
- a dataset $\boldsymbol{z} = \big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \big) \in (\mathcal{X} \times \mathcal{Y})^n$ **is available**

The Problem of Supervised Learning

Consider the following **supervised learning** scenario:

- three sets $\mathcal{X}$ (patterns), $\mathcal{Y}$ (labels) and $\mathcal{M} \subset \mathbb{R}^d$ (models), with $d \in \mathbb{N}$.
- a function $f : \mathcal{M} \times \mathcal{X} \to \mathcal{Y}$ (explicit expression **is known**)

Statistical Assumptions

Two random variables X and Y satisfy

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

for some specific model $\boldsymbol{\theta}^\star$ (optimal model or hypothesis).

Objective: Model Selection

Given a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, find the model $\boldsymbol{\theta}^\star$ in (1)

The Problem of Supervised Learning: **Empirical Risk Minimization**

- Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$ be a **risk (or loss or cost)** function.

Risk

Given a data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the model $\boldsymbol{\theta} \in \mathcal{M}$ induces the **risk** $\ell\left(f(\boldsymbol{\theta}, x), y\right)$.

Empirical Risk

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(f(\boldsymbol{\theta}, x_i), y_i\right). \tag{2}$$

The Problem of Supervised Learning: **Empirical Risk Minimization**

- Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$ be a **risk (or loss or cost)** function.

Risk

Given a data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the model $\boldsymbol{\theta} \in \mathcal{M}$ induces the **risk** $\ell\left(f(\boldsymbol{\theta}, x), y\right)$.

Empirical Risk

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(f(\boldsymbol{\theta}, x_i), y_i\right). \tag{2}$$

Problem Formulation: Empirical Risk Minimization (ERM) [Bishop, 2006]

Given the dataset $\boldsymbol{z}$,

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

The Problem of Supervised Learning: **Empirical Risk Minimization**

Problem Formulation: Empirical Risk Minimization (ERM) [Bishop, 2006]

Given the dataset $\boldsymbol{z}$,

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}). \tag{3}$$

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

$$\mathcal{T}(\boldsymbol{z}) \triangleq \arg\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}). \tag{4}$$

The Problem of Supervised Learning: **Empirical Risk Minimization**

Problem Formulation: Empirical Risk Minimization (ERM) [Bishop, 2006]

Given the dataset $\boldsymbol{z}$,

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathrm{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) . \tag{3}$$

Strong connection with other problems:

- M-Estimation [Stefanski and Boos, 2002]
- minimum contrast estimation [Massart, 2007, Birge and Massart, 1993]
- sample average approximation [Kleywegt et al., 2002]

Appears in: machine learning [Chaudhuri et al., 2011], statistical physics, statistics, operations research, decision making, game theory, information theory, stochastic optimization, ...

Methods for Solving ERM problems

- Gradient-based methods
  - Gradient Descent [Cauchy, 1847]
  - Stochastic Gradient Descent [Robbins and Monro, 1951]
  - Stochastic Mirror Descent
  - AdaGrad, Adam, RMSprop, ... [Kingma and Ba, 2014]
  - State of the art in [Bottou et al., 2018, Xin et al., 2020]
- Gradient-free methods
  - Particle Swarm Optimization [Kennedy and Eberhart, 1995]
  - Surrogate optimization
  - **Simulated Annealing** (ERM with Regularization by entropy)

Methods for Solving ERM problems: **Gradient-Based Methods**

- L. A. Cauchy, **"Méthode générale pour la résolution des systèmes d'équations simultanées"**. Compte Rendu à l'Académie des Sciences, Octobre 18, 1847.
- **Limitations** [Jin et al., 2017]:
  - **Local Optimizers** might be accepted as global optimizers
  - **Slow Convergence Rates** even for large datasets
  - **Learning Rate** trades off ability to overlook local minimizers and runtime
  - **Initialization Point** determines global performance
  - **Gradient-friendly** loss functions

Methods for Solving ERM problems: **Gradient-Free Methods**

Focus is on **simulated annealing** methods:

- A **probability measure** $P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$ on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ **conditioned** on a dataset $\boldsymbol{z}$

Methods for Solving ERM problems: **Gradient-Free Methods**

Focus is on **simulated annealing** methods:

- A **probability measure** $P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$ on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ **conditioned** on a dataset $\boldsymbol{z}$
- **What to do** with such a probability measure $P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$?
  - **Sampling** models from $P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$
  - **Maximum à posteriori (MAP)** models: $\hat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in\mathcal{M}} P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})$
  - **Expected** model: $\hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta} \mathrm{d}P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})$

Methods for Solving ERM problems: **Gradient-Free Methods**

Focus is on **simulated annealing** methods:

- A **probability measure** $P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$ on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ **conditioned** on a dataset $\boldsymbol{z}$
- **What to do** with such a probability measure $P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$?
  - **Sampling** models from $P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$
  - **Maximum à posteriori (MAP)** models: $\hat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in\mathcal{M}} P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})$
  - **Expected** model: $\hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta} \mathrm{d}P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})$
- **How to build such probability measures ?**
  - Statistical **assumptions on the datasets** for some specific loss functions
  - Bayesian methods [Guedj, 2019]
  - ERM with **Regularization by Relative Entropy** [Catoni, 2007, Zhang, 2006, Asadi et al., 2018, Jaakkola et al., 1999, Xu and Raginsky, 2017]

Table of Contents

ERM with Relative Entropy Regularization (ERM-RER)

Definition (Expected Empirical Risk)

Given a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $\mathsf{R}_{\boldsymbol{z}} : \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to [0, +\infty)$ be such that for all σ-finite measures $P \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}). \tag{4}$$

When P is a probability measure, the **expected empirical risk** induced by P is $\mathsf{R}_{\boldsymbol{z}}(P)$.

ERM with Relative Entropy Regularization (ERM-RER)

Definition (Expected Empirical Risk)

Given a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $\mathrm{R}_{\boldsymbol{z}} : \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to [0, +\infty)$ be such that for all σ-finite measures $P \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that

$$\mathrm{R}_{\boldsymbol{z}}(P) = \int \mathrm{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}). \tag{4}$$

When P is a probability measure, the **expected empirical risk** induced by P is $\mathrm{R}_{\boldsymbol{z}}(P)$.

Definition (Generalized Relative Entropy)

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

$$D(Q||P) \triangleq \int \frac{\mathrm{d}Q}{\mathrm{d}P}(x) \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(x)\right) \mathrm{d}P(x), \tag{5}$$

ERM with Relative Entropy Regularization (ERM-RER)

Definition (Expected Empirical Risk)

Given a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $\mathsf{R}_{\boldsymbol{z}} : \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to [0, +\infty)$ be such that for all σ-finite measures $P \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}). \tag{4}$$

When P is a probability measure, the **expected empirical risk** induced by P is $\mathsf{R}_{\boldsymbol{z}}(P)$.

Definition (ERM-RER Problem)

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:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(P \| Q). \tag{5}$$

ERM with Relative Entropy Regularization (ERM-RER)

Definition (ERM-RER Problem)

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:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{Z}}(P) + \lambda D(P \| Q). \tag{4}$$

Example 1: Assume that $\mathcal{M} = \mathbb{R}^d$ and Q is the **Lebesgue measure**:

$$\min_{g} \int_{\mathcal{M}} \mathsf{L}_{\boldsymbol{Z}}(\boldsymbol{\theta}) g(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} - \lambda H(g), \text{ with } H \text{ the entropy function.} \tag{5}$$

The solution is for all $\boldsymbol{\theta} \in \mathcal{M}$

$$g^{(\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int_{\mathbb{R}^d} \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}\boldsymbol{\nu}}. \tag{6}$$

ERM with Relative Entropy Regularization (ERM-RER)

Definition (ERM-RER Problem)

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:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{Z}}(P) + \lambda D(P \| Q). \tag{4}$$

Example 2: Assume that $\mathcal{M}$ is countable and Q is the **counting measure**:

$$\min_{p} \sum_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{Z}}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}) - \lambda H(p), \text{ with } H \text{ the entropy function.} \tag{5}$$

The solution is for all $\boldsymbol{\theta} \in \mathcal{M}$

$$p^{(\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\sum_{\boldsymbol{\nu} \in \mathcal{M}} \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}. \tag{6}$$

ERM with Relative Entropy Regularization (ERM-RER)

Definition (ERM-RER Problem)

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:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(P \| Q). \tag{4}$$

Example 3: The set $\mathcal{M} \subset \mathbb{R}^d$ and the measure Q form a **probability space** $(\mathcal{M}, \mathscr{B}(\mathcal{M}), Q)$.

The solution is for all $\boldsymbol{\theta} \in \mathcal{M}$

$$P^{(\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int_{\mathcal{M}} \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\nu})}. \tag{5}$$

ERM with Relative Entropy Regularization (ERM-RER)

Definition (ERM-RER Problem)

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:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(P \| Q). \tag{4}$$

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

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp(t\, \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta})\right). \tag{5}$$

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

$$\mathcal{K}_{Q,\boldsymbol{z}} \triangleq \left\{\lambda \in (0, +\infty) : \; K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < +\infty\right\}. \tag{6}$$

ERM with Relative Entropy Regularization (ERM-RER)

Definition (ERM-RER Problem)

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:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{Z}}(P) + \lambda D(P \| Q). \tag{4}$$

Theorem

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

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{Z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right). \tag{5}$$

ERM with Relative Entropy Regularization (ERM-RER)

Definition (ERM-RER Problem)

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:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathrm{R}_{\boldsymbol{z}}(P) + \lambda D(P \| Q). \tag{4}$$

What is the motivation of this generalization?

- Some priors are not possible to be formed as 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

ERM-RER: Properties of the Solution

Lemma

For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, *under the assumption that*

$$Q\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = +\infty\}\right) = 0, \tag{5}$$

the σ*-finite measure* Q *and the probability measure* $P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$ *are mutually absolutely continuous.*

ERM-RER: Properties of the Solution

$$\delta^{\star}_{Q,\boldsymbol{z}} \triangleq \inf\left\{\delta \in [0,+\infty) : Q\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}\right) > 0\right\} \tag{6}$$

$$\mathcal{L}^{\star}_{Q,\boldsymbol{z}} = \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\} \tag{7}$$

Lemma

For all $\boldsymbol{\theta} \in \mathrm{supp}\mathrm{Q}$,

$$\lim_{\lambda \to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)} \mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}\}}. \tag{8}$$

ERM-RER: Properties of the Solution

$$\delta^{\star}_{Q,\boldsymbol{z}} \triangleq \inf\left\{\delta \in [0,+\infty) : Q\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}\right) > 0\right\} \tag{6}$$

$$\mathcal{L}^{\star}_{Q,\boldsymbol{z}} = \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\} \tag{7}$$

Lemma

The measure $P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$ and the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ satisfy

$$\lim_{\lambda \to 0^+} P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 1. \tag{8}$$

ERM-RER: **Coherent Measures**

Definition (Coherent Measures)

The σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ is said to be **coherent** if, for all $\delta > 0$, it holds that

$$Q\Big(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}\Big) > 0. \tag{9}$$

- If Q is not coherent, then $Q(\mathcal{T}(\boldsymbol{z})) = 0$.

Lemma

For all $\delta \in (0, +\infty)$ and for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$,

$$P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}) > 0, \tag{10}$$

if and only if the σ-finite measure Q is coherent.

ERM-RER: **Coherent Measures**

Definition (Coherent Measures)

The σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ is said to be **coherent** if, for all $\delta > 0$, it holds that

$$Q\Big(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{Z}}(\boldsymbol{\theta}) \leqslant \delta\}\Big) > 0. \tag{9}$$

- If Q is not coherent, then $Q(\mathcal{T}(\boldsymbol{z})) = 0$.

Lemma

The probability measure $P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$ and the set $\mathcal{T}(\boldsymbol{z})$ satisfy

$$\lim_{\lambda \to 0^+} P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{T}(\boldsymbol{z})) = 1, \tag{10}$$

if and only if the σ-finite measure Q is coherent.

Table of Contents

Probably Approximately Correct (PAC) Guarantees

Definition ((δ, ϵ)-Optimality)

Given a pair $(\delta, \epsilon) \in [0, +\infty) \times (0, 1)$, the probability measure $P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$, is said to be (δ, ϵ)-optimal, if

$$P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}\right) > 1 - \epsilon. \tag{11}$$

- For all $\delta > 0$, it holds that $\mathcal{T}(\boldsymbol{z}) \subset \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}$
- **Warning**:
  - For some Q and $\delta > 0$, $P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}\right) > 1 - \epsilon$ and $P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{T}(\boldsymbol{z})\right) = 0$.

Probably Approximately Correct (PAC) Guarantees

Definition ((δ, ϵ)-Optimality)

Given a pair $(\delta, \epsilon) \in [0, +\infty) \times (0, 1)$, the probability measure $P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$, is said to be (δ, ϵ)-optimal, if

$$P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}\right) > 1 - \epsilon. \tag{11}$$

Theorem

For all $(\delta, \epsilon) \in (\delta^{\star}_{Q,\boldsymbol{z}}, +\infty) \times (0, 1)$, there always exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ such that

$$P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}\right) > 1 - \epsilon, \tag{12}$$

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

Probably Approximately Correct (PAC) Guarantees

Definition ((δ, ϵ)-Optimality)

Given a pair $(\delta, \epsilon) \in [0, +\infty) \times (0, 1)$, the probability measure $P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$, is said to be (δ, ϵ)-optimal, if

$$P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}\right) > 1 - \epsilon. \tag{11}$$

Corollary

If the σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ is <u>coherent</u>, then for all $(\delta, \epsilon) \in (0, +\infty) \times (0, 1)$, there always exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ such that

$$P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}(\delta)\right) > 1 - \epsilon. \tag{12}$$

Statistical Properties of the Empirical Risk: **Cumulants**

Consider the random variable

$$\mathsf{L}_{\boldsymbol{z}}(\Theta), \text{ with } \Theta \sim P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}, \tag{13}$$

which models the empirical risk on the **training data** when $\Theta \sim P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$

Statistical Properties of the Empirical Risk: **Cumulants**

Consider the random variable

$$\mathsf{L}_{\boldsymbol{z}}(\Theta), \text{ with } \Theta \sim P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}, \tag{13}$$

which models the empirical risk on the **training data** when $\Theta \sim P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$

Lemma

For all $\lambda \in \operatorname{int}\mathcal{K}_{Q,\boldsymbol{z}}$,

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) = \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{14}$$

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)^2 \mathrm{d}P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \text{ and} \tag{15}$$

$$K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)^3 \mathrm{d}P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}). \tag{16}$$

Statistical Properties of the Empirical Risk: **Cumulants**

Consider the random variable

$$\mathsf{L}_{\boldsymbol{z}}(\Theta), \text{ with } \Theta \sim P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}, \tag{13}$$

which models the empirical risk on the **training data** when $\Theta \sim P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$

Theorem

The expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right)$ *is* nondecreasing *with* $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$. *Moreover,*

$$\lim_{\lambda \to 0^+} \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right) = \delta^{\star}_{Q,\boldsymbol{z}}, \tag{14}$$

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

Statistical Properties of the Empirical Risk: **(Trivial) Example**

Example

- Let $\mathcal{A} \subset \mathcal{M}$ and $\mathcal{M} \setminus \mathcal{A}$ be nonnegligible w.r.t. Q.
- For all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{A} \\ 1 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{M} \setminus \mathcal{A}. \end{cases} \tag{15}$$

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{\exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))}{Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))};$$

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{Q(\mathcal{A})(1 - Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)}{\left(Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))\right)^2}.$$

Statistical Properties of the Empirical Risk: **(Trivial) Example**

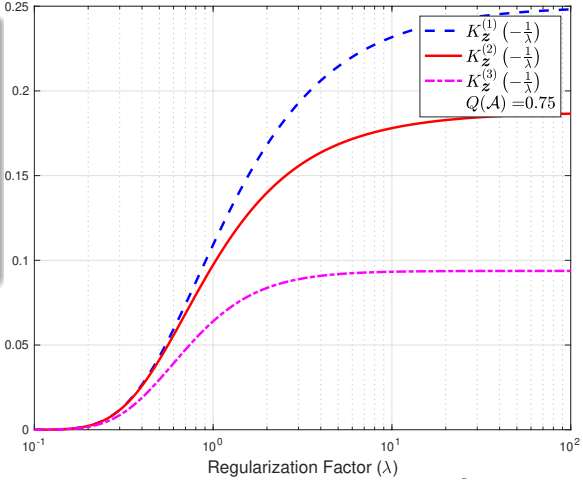
Example

- Let $\mathcal{A} \subset \mathcal{M}$ and $\mathcal{M} \setminus \mathcal{A}$ be nonnegligible w.r.t. Q.
- For all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{A} \\ 1 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{M} \setminus \mathcal{A}. \end{cases} \tag{15}$$

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{\exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))}{Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))};$$

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{Q(\mathcal{A})(1 - Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)}{\left(Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))\right)^2}$$



Statistical Properties of the Empirical Risk: **(Trivial) Example**

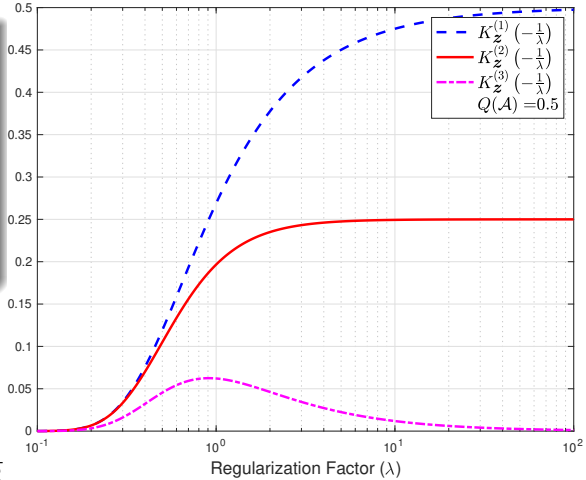
Example

- Let $\mathcal{A} \subset \mathcal{M}$ and $\mathcal{M} \setminus \mathcal{A}$ be nonnegligible w.r.t. Q.
- For all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{A} \\ 1 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{M} \setminus \mathcal{A}. \end{cases} \tag{15}$$

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{\exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))}{Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))};$$

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{Q(\mathcal{A})(1 - Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)}{\left(Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))\right)^2}$$



Statistical Properties of the Empirical Risk: **(Trivial) Example**

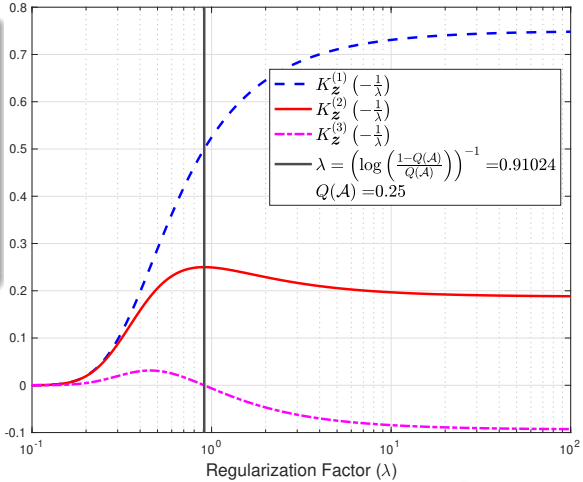
Example

- Let $\mathcal{A} \subset \mathcal{M}$ and $\mathcal{M} \setminus \mathcal{A}$ be nonnegligible w.r.t. Q.
- For all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{A} \\ 1 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{M} \setminus \mathcal{A}. \end{cases} \tag{15}$$

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{\exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))}{Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))};$$

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{Q(\mathcal{A})(1 - Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)}{\left(Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))\right)^2}$$



Statistical Properties of the Empirical Risk: **Sub-Gaussianity**

Consider the random variable

$$\mathsf{L}_{\boldsymbol{z}}(\Theta), \text{ with } \Theta \sim P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}, \tag{16}$$

whose **cumulant generating function** is $J_{\boldsymbol{z},\lambda}(t) = \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{u})\right)$.

Statistical Properties of the Empirical Risk: **Sub-Gaussianity**

Consider the random variable

$$\mathsf{L}_{\boldsymbol{z}}(\Theta), \text{ with } \Theta \sim P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}, \tag{16}$$

whose **cumulant generating function** is $J_{\boldsymbol{z},\lambda}(t) = \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{u})\right)$.

Lemma

Given a real $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the cumulant generating function $J_{\boldsymbol{z},\lambda}$ verifies for all $t \in \left(-\infty, \frac{1}{\lambda}\right)$,

$$J_{\boldsymbol{z},\lambda}(t) = K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < +\infty. \tag{17}$$

Statistical Properties of the Empirical Risk: **Sub-Gaussianity**

Consider the random variable

$$\mathsf{L}_{\boldsymbol{z}}(\Theta), \text{ with } \Theta \sim P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}, \tag{16}$$

whose **cumulant generating function** is $J_{\boldsymbol{z},\lambda}(t) = \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{u})\right)$.

Theorem

For all $\alpha \in \left(-\infty, \frac{1}{\lambda}\right)$,

$$J_{\boldsymbol{z},\lambda}(\alpha) \leqslant \alpha K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{1}{2}\alpha^2 B^2_{\boldsymbol{z}} \tag{17}$$

where, the constant $B_{\boldsymbol{z}} > 0$ *satisfies*

$$B^2_{\boldsymbol{z}} = \sup_{\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}} K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right). \tag{18}$$

Statistical Properties of the Empirical Risk: **Sub-Gaussianity**

Consider the random variable

$$\mathsf{L}_{\boldsymbol{z}}(\Theta), \text{ with } \Theta \sim P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}, \tag{16}$$

whose **cumulant generating function** is $J_{\boldsymbol{z},\lambda}(t) = \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{u})\right)$.

Sub-Gaussianity

The random variable $\mathsf{L}_{\boldsymbol{z}}(\Theta)$ is sub-Gaussian, with sub-Gaussian parameter $B_{\boldsymbol{z}}$.

Sensitivity: **Definition**

Definition (Sensitivity)

Let $\mathsf{S}_{Q,\lambda} : (\mathcal{X} \times \mathcal{Y})^n \times \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to (-\infty, +\infty]$ be a function such that

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \begin{cases} \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right) & \text{if } \lambda \in \mathcal{K}_{Q,\boldsymbol{z}} \\ +\infty & \text{otherwise.} \end{cases} \tag{17}$$

Sensitivity: **Definition**

Definition (Sensitivity)

Let $\mathsf{S}_{Q,\lambda} : (\mathcal{X} \times \mathcal{Y})^n \times \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to (-\infty, +\infty]$ be a function such that

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \begin{cases} \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right) & \text{if } \lambda \in \mathcal{K}_{Q,\boldsymbol{z}} \\ +\infty & \text{otherwise.} \end{cases} \tag{17}$$

Theorem (Explicit Expression for Sensitivity)

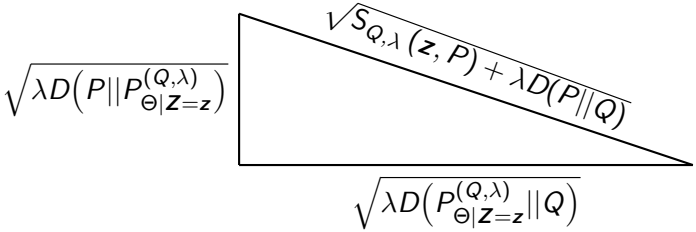
$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \lambda\Big(D\big(P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}} || Q\big) + D\big(P || P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\big) - D(P||Q)\Big)$$

Sensitivity: **Definition**

Definition (Sensitivity)

Let $\mathsf{S}_{Q,\lambda} : (\mathcal{X} \times \mathcal{Y})^n \times \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to (-\infty, +\infty]$ be a function such that

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \begin{cases} \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right) & \text{if } \lambda \in \mathcal{K}_{Q,\boldsymbol{z}} \\ +\infty & \text{otherwise.} \end{cases} \tag{17}$$



Sensitivity: **Dataset-dependent Bounds**

Theorem

Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, it holds that, for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ and for all probability measures $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$,

$$\left|\mathrm{R}_{\boldsymbol{z}}(P) - \mathrm{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right)\right| \leqslant \sqrt{2B^2_{Q,\boldsymbol{z}} D\left(P \| P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right)}, \tag{18}$$

where the constant $B_{Q,\boldsymbol{z}} \in [0, +\infty)$ is

$$B^2_{Q,\boldsymbol{z}} = \sup_{\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}} K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right). \tag{19}$$

Sensitivity: **Dataset-independent Bounds**

Corollary

Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, for all $\lambda \in \mathcal{K}_Q$, and for all probability measures $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that

$$\int \left| \mathrm{R}_{\boldsymbol{z}}(P) - \mathrm{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right) \right| \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \leqslant \int \sqrt{2B^2_{Q,\boldsymbol{z}} D\left(P \| P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right)} \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{20}$$

where the constant $B_{Q,\boldsymbol{z}} \in [0, +\infty)$ is

$$B^2_{Q,\boldsymbol{z}} = \sup_{\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}} K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right). \tag{21}$$

Sensitivity: **Dataset-independent Bounds**

Theorem

Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, for all $\lambda \in \mathcal{K}_Q$ and for all probability measures $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that

$$\int \left| \mathrm{R}_{\boldsymbol{z}}(P) - \mathrm{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right) \right| \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \leqslant \sqrt{2B_Q^2 \int D\left(P \| P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{u}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u})},$$

where the constant B_Q satisfies

$$B_Q^2 = \sup_{\boldsymbol{z} \in \mathrm{supp}\mathrm{P}_{\mathrm{Z}}} B_{Q,\boldsymbol{z}}^2. \tag{20}$$

Sensitivity: **Dataset-independent Bounds**

Theorem

Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, for all $\lambda \in \mathcal{K}_Q$, it holds that

$$\int \left| \mathrm{R}_{\boldsymbol{z}}\left(P_{\Theta}^{(Q,\lambda)}\right) - \mathrm{R}_{\boldsymbol{z}}\left(P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) \right| \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \leqslant \sqrt{2B_Q^2 \mathrm{L}(\boldsymbol{Z};\Theta)},$$

where the probability measure $P_{\Theta}^{(Q,\lambda)}$ is such that for all $\mathcal{A} \in \mathscr{B}(\mathcal{M})$,

$$P_{\Theta}^{(Q,\lambda)}(\mathcal{A}) = \int P_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}(\mathcal{A}) \, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \tag{20}$$

and the constant B_Q satisfies

$$B_Q^2 = \sup_{\boldsymbol{z} \in \mathrm{supp} \mathrm{P}_{\mathrm{Z}}} B_{Q,\boldsymbol{z}}^2. \tag{21}$$

Universality

Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, and a nonnegative real $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, consider the following optimization problem

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}P(\boldsymbol{\theta}), \tag{22a}$$

$$\text{subject to:} \quad D\left(P \| P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right) \leqslant c, \text{ with } c > 0. \tag{22b}$$

Theorem

The solution to the optimization problem in (22) is a probability measure $P^{(Q,\omega)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}$ satisfying for all $\boldsymbol{\theta} \in \mathrm{suppP}$,

$$\frac{\mathrm{d}P^{(Q,\omega)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\omega}\right) - \frac{1}{\omega}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right), \tag{23}$$

with $\omega \in (0, \lambda]$ such that $D\left(P^{(Q,\omega)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(Q,\lambda)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}}\right) = c$.

Table of Contents

Generalization Capabilities: **Dataset Aggregation**

- **Three datasets.** For all $i \in \{0, 1, 2\}$

$$\boldsymbol{z}_i = ((x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \ldots, (x_{i,n_i}, y_{i,n_i})) \in (\mathcal{X} \times \mathcal{Y})^{n_i}, \text{ with } n_0 = n_1 + n_2 \quad (24)$$

- **Constituent** datasets: $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$
- **Aggregated** dataset: $\boldsymbol{z}_0 = (\boldsymbol{z}_1, \boldsymbol{z}_2)$
- Interpretation 1: Training - Validation
- Interpretation 2: Decentralized learning
- Interpretation 3: Different data acquision systems

Generalization Capabilities: **Dataset Aggregation**

- **Three datasets.** For all $i \in \{0, 1, 2\}$

$$\boldsymbol{z}_i = ((x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \ldots, (x_{i,n_i}, y_{i,n_i})) \in (\mathcal{X} \times \mathcal{Y})^{n_i}, \text{ with } n_0 = n_1 + n_2 \quad (24)$$

- **Constituent** datasets: $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$
- **Aggregated** dataset: $\boldsymbol{z}_0 = (\boldsymbol{z}_1, \boldsymbol{z}_2)$
- Interpretation 1: Training - Validation
- Interpretation 2: Decentralized learning
- Interpretation 3: Different data acquision systems
- Each dataset induces **a different ERM-RER problem**:

$$\min_{P \in \triangle_{Q_i}(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathrm{R}_{\boldsymbol{z}_i}(P) + \lambda_i D(P \| Q_i). \quad (25)$$

Generalization Capabilities: **Cross-Validation**

- **Three datasets.** For all $i \in \{0,1,2\}$

$$\boldsymbol{z}_i = ((x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \ldots, (x_{i,n_i}, y_{i,n_i})) \in (\mathcal{X} \times \mathcal{Y})^{n_i}, \text{ with } n_0 = n_1 + n_2 \quad (26)$$

Theorem (Cross-Validation Error)

For all $i \in \{1,2\}$ and $j \in \{1,2\} \setminus \{i\}$:

$$\begin{aligned} &\mathrm{R}_{\boldsymbol{z}_i}\left(P^{(Q_j,\lambda_j)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_j}\right) - \mathrm{R}_{\boldsymbol{z}_i}\left(P^{(Q_i,\lambda_i)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_i}\right) \\ =&\lambda_i\left(D\left(P^{(Q_j,\lambda_j)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_j}||P^{(Q_i,\lambda_i)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_i}\right) + D\left(P^{(Q_i,\lambda_i)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_i}||Q_i\right) - D\left(P^{(Q_j,\lambda_j)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_j}||Q_i\right)\right). \end{aligned} \quad (27)$$

Generalization Capabilities: **Quasi-Generalization Error**

- **Three datasets.** For all $i \in \{0,1,2\}$

$$\boldsymbol{z}_i = ((x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \ldots, (x_{i,n_i}, y_{i,n_i})) \in (\mathcal{X} \times \mathcal{Y})^{n_i}, \text{ with } n_0 = n_1 + n_2 \quad (28)$$

Theorem (Quasi-Generalization Error)

For all $i \in \{0,1,2\}$,

$$\begin{aligned} &\mathrm{R}_{\boldsymbol{z}_0}\left(P^{(Q_i,\lambda_i)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_i}\right) - \mathrm{R}_{\boldsymbol{z}_0}\left(P^{(Q_0,\lambda_0)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_0}\right) \\ =&\lambda_0\left(D\Big(P^{(Q_i,\lambda_i)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_i}||P^{(Q_0,\lambda_0)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_0}\Big) + D\Big(P^{(Q_0,\lambda_0)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_0}||Q_0\Big) - D\Big(P^{(Q_i,\lambda_i)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_i}||Q_0\Big)\right). \end{aligned} \quad (29)$$

Generalization Capabilities:
Homogeneous Priors and Proportional Regularization

- **Three datasets.** For all $i \in \{0,1,2\}$

$$\boldsymbol{z}_i = ((x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \ldots, (x_{i,n_i}, y_{i,n_i})) \in (\mathcal{X} \times \mathcal{Y})^{n_i}, \text{ with } n_0 = n_1 + n_2 \quad (30)$$

- For all $i \in \{0,1,2\}$ and for all $\mathcal{A} \in \mathscr{B}(\mathcal{M})$, $Q(\mathcal{A}) = Q_i(\mathcal{A})$ (**Homogeneous Priors**).
- $\lambda_1 = \frac{n_0}{n_1}\lambda_0$ and $\lambda_2 = \frac{n_0}{n_2}\lambda_0$, (**Proportional Regularization**).

Lemma

For all $\boldsymbol{\theta} \in \mathcal{M}$, $\mathrm{L}_{\boldsymbol{z}_0}(\boldsymbol{\theta}) = \frac{n_1}{n_0}\mathrm{L}_{\boldsymbol{z}_1}(\boldsymbol{\theta}) + \frac{n_2}{n_0}\mathrm{L}_{\boldsymbol{z}_2}(\boldsymbol{\theta})$.

For all σ*-finite measures* $P \in \triangle(M, \mathscr{B}(M))$, $\mathrm{R}_{\boldsymbol{z}_0}(P) = \frac{n_1}{n_0}\mathrm{R}_{\boldsymbol{z}_1}(P) + \frac{n_2}{n_0}\mathrm{R}_{\boldsymbol{z}_2}(P)$.

Generalization Capabilities:
Homogeneous Priors and Proportional Regularization

- **Three datasets.** For all $i \in \{0,1,2\}$

$$\boldsymbol{z}_i = \left((x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \ldots, (x_{i,n_i}, y_{i,n_i}) \right) \in (\mathcal{X} \times \mathcal{Y})^{n_i}, \text{ with } n_0 = n_1 + n_2 \quad (30)$$

- For all $i \in \{0,1,2\}$ and for all $\mathcal{A} \in \mathscr{B}(\mathcal{M})$, $Q(\mathcal{A}) = Q_i(\mathcal{A})$ (**Homogeneous Priors**).
- $\lambda_1 = \frac{n_0}{n_1}\lambda_0$ and $\lambda_2 = \frac{n_0}{n_2}\lambda_0$, (**Proportional Regularization**).

Theorem (Cross Generalization Error)

For all $i \in \{1,2\}$,

$$\mathrm{R}_{\boldsymbol{z}_0}\left(P^{(Q,\lambda_1)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathrm{R}_{\boldsymbol{z}_0}\left(P^{(Q,\lambda_2)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_2}\right) =$$

$$\lambda_0 \left(D\left(P^{(Q,\lambda_1)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_1} || P^{(Q,\lambda_2)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_2}\right) + 2D\left(P^{(Q,\lambda_1)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_1} || Q\right) - D\left(P^{(Q,\lambda_2)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_2} || P^{(Q,\lambda_1)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - 2D\left(P^{(Q,\lambda_2)}_{\Theta|\boldsymbol{Z}=\boldsymbol{z}_2} || Q\right) \right).$$

Table of Contents

Conclusions and Final Remarks

- ERM-RER: An alternative for **gradient-free** machine learning
- No assumption on the **distribution of the datasets**
- No requirements on the **loss function** (other than measurability)
- **Probably approximately correct** (PAC) guarantee
- Empirical risk is a **sub-Gaussian** random variable
- Sensitivity as an **alternative performance measure**

Conclusions and Final Remarks

- ERM-RER: An alternative for **gradient-free** machine learning
- No assumption on the **distribution of the datasets**
- No requirements on the **loss function** (other than measurability)
- **Probably approximately correct** (PAC) guarantee
- Empirical risk is a **sub-Gaussian** random variable
- Sensitivity as an **alternative performance measure**
- Further Research – Open Questions
  - Algorithms: Efficient construction of the Gibbs measure ?
  - Algorithms: What to do with new data for updating the Gibbs measure ?
  - Decentralized Learning: Message passing between learners ?
  - Federated Learning: Data-Aggregation ?

Thank you for your attention. Questions ?

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.

Birge, L. and Massart, P. (1993).

Rates of convergence for minimum contrast estimators.

Probability Theory and Related Fields, 97:113–150.

Bibliography II

Bishop, C. M. (2006).

Pattern Recognition and Machine Learning.

Springer-Verlag, Berlin, Heidelberg, first edition.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018).

Optimization methods for large-scale machine learning.

SIAM Review, 60(2):223–311.

Bibliography III

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.

Cauchy, A.-L. (1847).

Méthode générale pour la résolution des systèmes d'équations simultanées.

In *Compte Rendu à l'Académie des Sciences*, volume XXV, page 536, Paris, France.

Bibliography IV

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011).

Differentially private empirical risk minimization.

Journal of Machine Learning Research, 12(3):1069–1109.

Guedj, B. (2019).

A primer on PAC-Bayesian learning.

In *Tutorials of the International Conference on Machine Learning (ICML)*.

Jaakkola, T., Meila, M., and Jebara, T. (1999).

Maximum entropy discrimination.

In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada.

Bibliography V

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. (2017).

How to escape saddle points efficiently.

In *International Conference on Machine Learning*, pages 1724–1732. PMLR.

Kennedy, J. and Eberhart, R. (1995).

Particle swarm optimization.

In *Proc. of International Conference on Neural Networks*, volume 4, pages 1942–1948 vol.4.

Kingma, D. P. and Ba, J. (2014).

Adam: A method for stochastic optimization.

arXiv preprint arXiv:1412.6980.

Bibliography VI

Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. (2002).

The sample average approximation method for stochastic discrete optimization.

SIAM Journal on Optimization, 12(2):479–502.

Massart, P. (2007).

Concentration inequalities and model selection.

Springer, New York, NY, first edition.

Robbins, H. and Monro, S. (1951).

A stochastic approximation method.

The Annals of Mathematical Statistics, 22(3):400–407.

Bibliography VII

Stefanski, L. A. and Boos, D. D. (2002).

The calculus of M-estimation.

The American Statistician, 56(1):29–38.

Xin, R., Kar, S., and Khan, U. A. (2020).

Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence.

IEEE Signal Processing Magazine, 37(3):102–113.

Bibliography VIII

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.