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:


Table of Contents

1. The Problem of Supervised Learning
2. Empirical Risk Minimization with Relative Entropy Regularization
3. Probably Approximately Correct (PAC) Guarantees
4. Generalization Capabilities
5. Conclusions and Final Remarks
Table of Contents

1 The Problem of Supervised Learning

2 Empirical Risk Minimization with Relative Entropy Regularization

3 Probably Approximately Correct (PAC) Guarantees

4 Generalization Capabilities

5 Conclusions and Final Remarks
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 $f: \mathcal{M} \times \mathcal{X} \rightarrow \mathcal{Y}$ (explicit expression **is known**)

### Statistical Assumptions

Two random variables $X$ and $Y$ satisfy

$$Y = f(\theta^*, X),$$

for some specific model $\theta^*$ (optimal model or hypothesis).

- model $\theta^*$ is **unknown**
- a dataset $z = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ 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 \( f : \mathcal{M} \times \mathcal{X} \rightarrow \mathcal{Y} \) (explicit expression is known)

**Statistical Assumptions**

Two random variables \( X \) and \( Y \) satisfy

\[
Y = f(\theta^*, X),
\]

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

**Objective: Model Selection**

Given a dataset \( z \in (\mathcal{X} \times \mathcal{Y})^n \), find the model \( \theta^* \) in (1)
Let \( \ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, +\infty) \) be a **risk (or loss or cost)** function.

**Risk**

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

**Empirical Risk**

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

\[
L_z (\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell (f(\theta, x_i), y_i). \quad (2)
\]
Let $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, +\infty)$ be a risk (or loss or cost) function.

**Risk**

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

**Empirical Risk**

Given a dataset $z = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, 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) [Bishop, 2006]**

Given the dataset $z$,

$$\min_{\theta \in \mathcal{M}} L_z(\theta).$$
The Problem of Supervised Learning: **Empirical Risk Minimization**

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

Given the dataset $z$,

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

(3)

Set of *solutions to the ERM problem*:

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

(4)
The Problem of Supervised Learning: **Empirical Risk Minimization**

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

Given the dataset \( z \),

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

(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**


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

What to do with such a probability measure $P_{\Theta|Z=z}$?

- **Sampling** models from $P_{\Theta|Z=z}$
- Maximum `a posteriori (MAP) models: $\hat{\theta} \in \text{arg max}_{\theta \in \mathcal{M}} P_{\Theta|Z=z}(\theta)$
- **Expected model:** $\hat{\theta} = \int_{\theta} \theta \, dP_{\Theta|Z=z}(\theta)$

How to build such probability measures?

- Statistical assumptions on the datasets for some specific loss functions
- **Bayesian methods** [Guedj, 2019]
Methods for Solving ERM problems: Gradient-Free Methods

Focus is on **simulated annealing** methods:

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

- **What to do** with such a probability measure $P_{\Theta|Z=z}$?
  - **Sampling** models from $P_{\Theta|Z=z}$
  - **Maximum à posteriori (MAP)** models: $\hat{\theta} \in \arg\max_{\theta \in \mathcal{M}} P_{\Theta|Z=z}(\theta)$
  - **Expected** model: $\hat{\theta} = \int \theta dP_{\Theta|Z=z}(\theta)$

How to build such probability measures?
- **Statistical assumptions on the datasets for some specific loss functions**
- **Bayesian methods** [Guedj, 2019]
Methods for Solving ERM problems: **Gradient-Free Methods**

Focus is on simulated annealing methods:

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

- **What to do** with such a probability measure $P_{\Theta|Z=z}$?
  - **Sampling** models from $P_{\Theta|Z=z}$
  - **Maximum à posteriori (MAP)** models: $\hat{\theta} \in \arg\max_{\theta \in \mathcal{M}} P_{\Theta|Z=z}(\theta)$
  - **Expected** model: $\hat{\theta} = \int \theta dP_{\Theta|Z=z}(\theta)$

- **How to build such probability measures?**
  - Statistical **assumptions on the datasets** for some specific loss functions
  - Bayesian methods [Guedj, 2019]
Table of Contents

1. The Problem of Supervised Learning

2. Empirical Risk Minimization with Relative Entropy Regularization

3. Probably Approximately Correct (PAC) Guarantees

4. Generalization Capabilities

5. Conclusions and Final Remarks
ERM with Relative Entropy Regularization (ERM-RER)

Definition (Expected Empirical Risk)

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

\[
R_z (P) = \int L_z (\theta) \, dP(\theta).
\]  

(4)

When \( P \) is a probability measure, the expected empirical risk induced by \( P \) is \( R_z (P) \).

When \( P \) is a probability measure, the expected empirical risk induced by \( P \) is \( R_z (P) \).
ERM with Relative Entropy Regularization (ERM-RER)

**Definition (Expected Empirical Risk)**

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

\[
R_z(P) = \int L_z(\theta) \, dP(\theta).
\]

(4)

When \( P \) is a probability measure, the **expected empirical risk** induced by \( P \) is \( R_z(P) \).

**Definition (Generalized Relative Entropy)**

Given two \( \sigma \)-finite measures \( P \) and \( Q \) on the same measurable space, such that \( Q \ll P \),

\[
D(Q||P) \triangleq \int \frac{dQ}{dP}(x) \log \left( \frac{dQ}{dP}(x) \right) \, dP(x),
\]

(5)
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 $\sigma$-finite measures $P \in \Delta (\mathcal{M}, \mathcal{B}(\mathcal{M}))$, it holds that

$$R_{\mathbf{z}} (P) = \int L_{\mathbf{z}} (\theta) \, dP(\theta).$$  \hspace{1cm} (4)

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

Definition (ERM-RER Problem)

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

$$\min_{P \in \Delta_Q (\mathcal{M}, \mathcal{B}(\mathcal{M}))} R_{\mathbf{z}} (P) + \lambda D (P \parallel Q).$$  \hspace{1cm} (5)
ERM with Relative Entropy Regularization (ERM-RER)

**Definition (ERM-RER Problem)**

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

$$\min_{P \in \Delta_Q(\mathcal{M}, \mathcal{B}(\mathcal{M}))} R_z(P) + \lambda D(P \parallel Q).$$  \hspace{1cm} (4)

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

$$\min_g \int_{\mathcal{M}} L_z(\theta) g(\theta) \, d\theta - \lambda H(g), \text{ with } H \text{ the entropy function.}$$ \hspace{1cm} (5)

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

$$g^{(\lambda)}_{\theta|Z=z}(\theta) = \frac{\exp \left( - \frac{L_z(\theta)}{\lambda} \right)}{\int_{\mathbb{R}^d} \exp \left( - \frac{L_z(\nu)}{\lambda} \right) \, d\nu}. \hspace{1cm} (6)$$
Definition (ERM-RER Problem)

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

$$\min_{P \in \Delta Q(\mathcal{M}, \mathcal{B}(\mathcal{M}))} R_z(P) + \lambda D(P \parallel Q).$$  \hfill (4)

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

$$\min_p \sum_{\theta \in \mathcal{M}} L_z(\theta) p(\theta) - \lambda H(p), \text{ with } H \text{ the entropy function}. \hfill (5)$$

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

$$p_{\theta|Z=z}^{(\lambda)}(\theta) = \frac{\exp \left( - \frac{L_z(\theta)}{\lambda} \right)}{\sum_{\nu \in \mathcal{M}} \exp \left( - \frac{L_z(\nu)}{\lambda} \right)}. \hfill (6)$$
ERM with Relative Entropy Regularization (ERM-RER)

**Definition (ERM-RER Problem)**

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

$$\min_{P \in \triangle_Q (\mathcal{M}, \mathcal{B}(\mathcal{M}))} R_z (P) + \lambda D (P \parallel Q). \quad (4)$$

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

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

$$P_{\Theta|Z=z}^{(\lambda)} (\theta) = \frac{\exp \left( - \frac{L_z (\theta)}{\lambda} \right)}{\int_{\mathcal{M}} \exp \left( - \frac{L_z (\nu)}{\lambda} \right) dQ (\nu)}.$$  \hspace{1cm} (5)
ERM with Relative Entropy Regularization (ERM-RER)

**Definition (ERM-RER Problem)**

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

\[
\min_{P \in \Delta_Q(\mathcal{M}, \mathcal{B}(\mathcal{M}))} \mathcal{R}_z(P) + \lambda D(P \| Q).
\]

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

\[
\mathcal{K}_{Q,z} \triangleq \left\{ \lambda \in (0, +\infty) : \ K_{Q,z}\left(-\frac{1}{\lambda}\right) < +\infty \right\}.
\]
**Definition (ERM-RER Problem)**

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

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

(4)

**Theorem**

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

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

(5)
ERM with Relative Entropy Regularization (ERM-RER)

**Definition (ERM-RER Problem)**

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

$$\min_{P \in \Delta_Q(\mathcal{M}, \mathcal{B}(\mathcal{M}))} R_z(P) + \lambda D(P \| Q).$$  \hspace{1cm} (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.
Lemma

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

$$Q \left( \{ \theta \in \mathcal{M} : L_z(\theta) = +\infty \} \right) = 0,$$

the $\sigma$-finite measure $Q$ and the probability measure $P^{(Q,\lambda)}_{\Theta|Z=z}$ are mutually absolutely continuous.
ERM-RER: Properties of the Solution

\[ \delta_{Q,z}^* \overset{\Delta}{=} \inf \{ \delta \in [0, +\infty) : Q(\{ \theta \in \mathcal{M} : L_z(\theta) \leq \delta \}) > 0 \} \]  
\[ \mathcal{L}_{Q,z}^* = \{ \theta \in \mathcal{M} : L_z(\theta) = \delta_{Q,z}^* \} \]

Lemma

For all \( \theta \in \text{supp}Q, \)

\[ \lim_{\lambda \to 0^+} \frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = \frac{1}{Q(\mathcal{L}_{Q,z}^*)} \mathbf{1}\{\theta \in \mathcal{L}_{Q,z}^*\}. \]
\[ \delta_{Q,z}^* \triangleq \inf \{ \delta \in [0, +\infty) : Q(\{\theta \in \mathcal{M} : L_z(\theta) \leq \delta\}) > 0\} \quad (6) \]

\[ \mathcal{L}_{Q,z}^* = \{ \theta \in \mathcal{M} : L_z(\theta) = \delta_{Q,z}^* \} \quad (7) \]

**Lemma**

The measure \( P_{\Theta|Z=z}^{(Q,\lambda)} \) and the set \( \mathcal{L}_{Q,z}^* \) satisfy

\[ \lim_{\lambda \to 0^+} P_{\Theta|Z=z}^{(Q,\lambda)} (\mathcal{L}_{Q,z}^*) = 1. \quad (8) \]
**Definition (Coherent Measures)**

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

$$Q\left(\{\theta \in \mathcal{M} : L_z(\theta) \leq \delta\}\right) > 0. \quad (9)$$

- If $Q$ is not coherent, then $Q(T(z)) = 0$.

**Lemma**

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

$$P_{\Theta|Z=z}^{(Q,\lambda)} (\{\theta \in \mathcal{M} : L_z(\theta) \leq \delta\}) > 0,$$

if and only if the $\sigma$-finite measure $Q$ is coherent. \quad (10)
Definition (Coherent Measures)

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

$$Q\left(\{\theta \in \mathcal{M} : L_z(\theta) \leq \delta\}\right) > 0.$$  \hspace{1cm} (9)

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

Lemma

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

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

if and only if the σ-finite measure $Q$ is coherent.
# Table of Contents

1. The Problem of Supervised Learning

2. Empirical Risk Minimization with Relative Entropy Regularization

3. Probably Approximately Correct (PAC) Guarantees

4. Generalization Capabilities

5. Conclusions and Final Remarks
Definition ((\(\delta, \epsilon\))-Optimality)

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

\[
P_{\Theta|Z=z}^{(Q, \lambda)} \left( \{ \theta \in \mathcal{M} : L_z(\theta) \leq \delta \} \right) > 1 - \epsilon. \tag{11}
\]

- For all \(\delta > 0\), it holds that \(T(z) \subset \{ \theta \in \mathcal{M} : L_z(\theta) \leq \delta \}\)

- **Warning:**
  - For some \(Q\) and \(\delta > 0\), \(P_{\Theta|Z=z}^{(Q, \lambda)} \left( \{ \theta \in \mathcal{M} : L_z(\theta) \leq \delta \} \right) > 1 - \epsilon\) and \(P_{\Theta|Z=z}^{(Q, \lambda)} (T(z)) = 0\).
**Definition ((\(\delta, \epsilon\))-Optimality)**

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

\[
P_{\Theta | Z = z}^{(Q, \lambda)} (\{\theta \in M : L_z(\theta) \leq \delta\}) > 1 - \epsilon.
\]

(11)

**Theorem**

For all \((\delta, \epsilon) \in (\delta^*_Q, z, +\infty) \times (0, 1)\), there always exists a \(\lambda \in K_{Q, z}\) such that

\[
P_{\Theta | Z = z}^{(Q, \lambda)} (\{\theta \in M : L_z(\theta) \leq \delta\}) > 1 - \epsilon,
\]

(12)

where \(\delta^*_Q, z \coloneqq \inf \{\delta \in [0, +\infty) : Q(L_z(\delta)) > 0\}\).
**Probably Approximately Correct (PAC) Guarantees**

**Definition (\((\delta, \epsilon)-Optimality\))**

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

\[
P_{\Theta|Z=z}^{(Q,\lambda)} (\{\theta \in \mathcal{M} : L_z(\theta) \leq \delta\}) > 1 - \epsilon.
\]

(11)

**Corollary**

*If the \(\sigma\)-finite measure \(Q \in \Delta (\mathcal{M}, \mathcal{B}(\mathcal{M}))\) is coherent, then for all \((\delta, \epsilon) \in (0, +\infty) \times (0, 1)\), there always exists a \(\lambda \in K_{Q,z}\) such that*

\[
P_{\Theta|Z=z}^{(Q,\lambda)} (\mathcal{L}_z(\delta)) > 1 - \epsilon.
\]

(12)
Consider the random variable

\[ L_z (\Theta), \quad \text{with} \quad \Theta \sim P_{\Theta|Z=z}^{(Q,\lambda)} \]  

which models the empirical risk on the training data when \( \Theta \sim P_{\Theta|Z=z}^{(Q,\lambda)} \)
Statistical Properties of the Empirical Risk: **Cumulants**

Consider the random variable

\[ L_z(\Theta), \text{ with } \Theta \sim P_{\Theta \mid Z=z}^{(Q,\lambda)}, \]  

(13)

which models the empirical risk on the **training data** when \( \Theta \sim P_{\Theta \mid Z=z}^{(Q,\lambda)} \)

**Lemma**

*For all \( \lambda \in \text{int} \mathcal{K}_{Q,z} \),*

\[
K_{Q,z}^{(1)} \left( -\frac{1}{\lambda} \right) = \int L_z(\theta) \, dP_{\Theta \mid Z=z}^{(Q,\lambda)}(\theta) = R_z \left( P_{\Theta \mid Z=z}^{(Q,\lambda)} \right),
\]

(14)

\[
K_{Q,z}^{(2)} \left( -\frac{1}{\lambda} \right) = \int \left( L_z(\theta) - K_{Q,z}^{(1)} \left( -\frac{1}{\lambda} \right) \right)^2 \, dP_{\Theta \mid Z=z}^{(Q,\lambda)}(\theta), \text{ and}
\]

(15)

\[
K_{Q,z}^{(3)} \left( -\frac{1}{\lambda} \right) = \int \left( L_z(\theta) - K_{Q,z}^{(1)} \left( -\frac{1}{\lambda} \right) \right)^3 \, dP_{\Theta \mid Z=z}^{(Q,\lambda)}(\theta).
\]

(16)
Consider the random variable

\[ L_z(\Theta), \text{ with } \Theta \sim P^{(Q, \lambda)}_{\Theta \mid Z = z}, \]  

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

**Theorem**

The expected empirical risk \( R_z \left( P^{(Q, \lambda)}_{\Theta \mid Z = z} \right) \) is nondecreasing with \( \lambda \in \mathcal{K}_{Q, z} \). Moreover,

\[
\lim_{\lambda \to 0^+} R_z \left( P^{(Q, \lambda)}_{\Theta \mid Z = z} \right) = \delta_{Q, z}^*,
\]

where \( \delta_{Q, z}^* \triangleq \inf \{ \delta \in [0, +\infty) : Q \left( \mathcal{L}_z(\delta) \right) > 0 \} \).
Statistical Properties of the Empirical Risk: (Trivial) Example

Example

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

$$L_z(\theta) = \begin{cases} 
0 & \text{if } \theta \in A \\
1 & \text{if } \theta \in M \setminus A.
\end{cases}$$

(15)

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

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

Let $\mathcal{A} \subset \mathcal{M}$ and $\mathcal{M} \setminus \mathcal{A}$ be nonnegligible w.r.t. $Q$.

For all $\theta \in \mathcal{M}$,

$$L_z(\theta) = \begin{cases} 0 & \text{if } \theta \in \mathcal{A} \\ 1 & \text{if } \theta \in \mathcal{M} \setminus \mathcal{A} \end{cases}$$

(15)

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

$$K_{Q,z}^{(2)} \left( \frac{-1}{\lambda} \right) = \frac{Q(\mathcal{A}) \left( 1 - Q(\mathcal{A}) \right) \exp \left( -\frac{1}{\lambda} \right)}{\left( Q(\mathcal{A}) \left( 1 - \exp \left( -\frac{1}{\lambda} \right) \left( 1 - Q(\mathcal{A}) \right) \right) \right)^2}.$$
Example

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

$$L_z(\theta) = \begin{cases} 
0 & \text{if } \theta \in A \\
1 & \text{if } \theta \in M \setminus A.
\end{cases} \quad (15)$$

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

\[
K^{(2)}_{Q,z}
\left(\frac{-1}{\lambda}\right) = \frac{Q(A) (1 - Q(A)) \exp\left(-\frac{1}{\lambda}\right)}{(Q(A) + \exp\left(-\frac{1}{\lambda}\right) (1 - Q(A)))^2}.
\]
Statistical Properties of the Empirical Risk: (Trivial) Example

Example

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

$$L_z(\theta) = \begin{cases} 
0 & \text{if } \theta \in A \\
1 & \text{if } \theta \in M \setminus A.
\end{cases}$$  \hspace{1cm} (15)$$

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

$$K_{Q,z}^{(2)} \left( -\frac{1}{\lambda} \right) = \frac{Q(A) (1 - Q(A)) \exp \left( -\frac{1}{\lambda} \right)}{(Q(A) + \exp \left( -\frac{1}{\lambda} \right) (1 - Q(A)))^2}.$$
Consider the random variable

\[ L_z(\Theta), \text{ with } \Theta \sim P_{\Theta|Z=z}^{(Q,\lambda)}, \]  

(16)

whose **cumulant generating function** is

\[ J_{z,\lambda}(t) = \log \left( \int \exp \left( t L_z(u) \right) dP_{\Theta|Z=z}^{(Q,\lambda)}(u) \right). \]
Statistical Properties of the Empirical Risk: **Sub-Gaussianity**

Consider the random variable

\[ L_z (\Theta) , \text{ with } \Theta \sim P^{(Q,\lambda)}_{\Theta | Z = z} , \tag{16} \]

whose **cumulant generating function** is

\[ J_{z,\lambda}(t) = \log \left( \int \exp \left( t L_z (u) \right) dP^{(Q,\lambda)}_{\Theta | Z = z}(u) \right) . \]

**Lemma**

*Given a real \( \lambda \in K_{Q,z} \), the cumulant generating function \( J_{z,\lambda} \) verifies for all \( t \in (-\infty, \frac{1}{\lambda}) \),

\[ J_{z,\lambda}(t) = K_{Q,z} \left( t - \frac{1}{\lambda} \right) - K_{Q,z} \left( -\frac{1}{\lambda} \right) < +\infty. \tag{17} \]
Statistical Properties of the Empirical Risk: **Sub-Gaussianity**

Consider the random variable

$$L_z(\Theta), \text{ with } \Theta \sim P_{\Theta|Z=z}^{(Q,\lambda)},$$

whose **cumulant generating function** is

$$J_{z,\lambda}(t) = \log \left( \int \exp \left( t \sum u \right) dP_{\Theta|Z=z}^{(Q,\lambda)}(u) \right).$$

**Theorem**

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

$$J_{z,\lambda}(\alpha) \leq \alpha K^{(1)}_{Q,z} \left( -\frac{1}{\lambda} \right) + \frac{1}{2} \alpha^2 B^2_z$$

where, the constant $B_z > 0$ satisfies

$$B^2_z = \sup_{\gamma \in K^{(2)}_{Q,z}} K^{(2)}_{Q,z} \left( -\frac{1}{\gamma} \right).$$
Consider the random variable

\[ L_z(\Theta), \text{ with } \Theta \sim P_{\Theta|Z=z}^{(Q,\lambda)}, \]  

whose cumulant generating function is

\[ J_{z,\lambda}(t) = \log \left( \int \exp \left( t L_z(u) \right) dP_{\Theta|Z=z}^{(Q,\lambda)}(u) \right). \]

Sub-Gaussianity

The random variable \( L_z(\Theta) \) is sub-Gaussian, with sub-Gaussian parameter \( B_z \).
Sensitivity: **Definition**

**Definition (Sensitivity)**

Let $S_{Q,\lambda} : (\mathcal{X} \times \mathcal{Y})^n \times \Delta Q (\mathcal{M}, \mathcal{B} (\mathcal{M})) \rightarrow (-\infty, +\infty]$ be a function such that

$$S_{Q,\lambda} (z, P) = \begin{cases} R_z (P) - R_z \left( P_{\theta|Z=z}^{(Q,\lambda)} \right) & \text{if } \lambda \in \mathcal{K}_{Q,z} \\ +\infty & \text{otherwise.} \end{cases}$$

(17)
Sensitivity: Definition

Definition (Sensitivity)

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

$$S_{Q,\lambda}(z, P) = \begin{cases} R_z(P) - R_z(P_{\Theta|Z=z}^{(Q,\lambda)}) & \text{if } \lambda \in \mathcal{K}_{Q,z} \\ +\infty & \text{otherwise.} \end{cases}$$

(17)

Theorem (Explicit Expression for Sensitivity)

$$S_{Q,\lambda}(z, P) = \lambda \left( D(P_{\Theta|Z=z}^{(Q,\lambda)} \parallel Q) + D(P \parallel P_{\Theta|Z=z}^{(Q,\lambda)}) - D(P \parallel Q) \right)$$
Let $S_{Q,\lambda} : (X \times Y)^n \times \Delta_Q (\mathcal{M}, \mathcal{B} (\mathcal{M})) \to (-\infty, +\infty]$ be a function such that

$$S_{Q,\lambda} (z, P) = \begin{cases} R_z (P) - R_z \left( P_{\Theta|Z=z}^{(Q,\lambda)} \right) & \text{if } \lambda \in \mathcal{K}_{Q,z} \\ +\infty & \text{otherwise.} \end{cases}$$

(17)
Sensitivity: **Dataset-dependent Bounds**

**Theorem**

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

$$\left| R_z (P) - R_z \left( P^{(Q,\lambda)}_{\Theta|Z=z} \right) \right| \leq \sqrt{2B_{Q,z}^2 D \left( P \parallel P^{(Q,\lambda)}_{\Theta|Z=z} \right)}, \quad (18)$$

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

$$B_{Q,z}^2 = \sup_{\gamma \in \mathcal{K}_{Q,z}} K_{Q,z}^{(2)} \left( -\frac{1}{\gamma} \right). \quad (19)$$
Corollary

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

$$\int \left| R_z (P) - R_z \left( P^{(Q,\lambda)}_\Theta | Z = z \right) \right| dP_z(z) \leq \int \sqrt{2B_{Q,z}^2 D \left( P \parallel P^{(Q,\lambda)}_\Theta | Z = z \right)} dP_z(z), \quad (20)$$

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

$$B_{Q,z}^2 = \sup_{\gamma \in K_{Q,z}} K_{Q,z}^{(2)} \left( -\frac{1}{\gamma} \right). \quad (21)$$
Theorem

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

$$\int \left| R_z (P) - R_z (P_{\Theta | Z = z}^{(Q, \lambda)}) \right| \, dP_Z(z) \leq \sqrt{2B_Q^2 \int D (P \parallel P_{\Theta | Z = u}^{(Q, \lambda)}) \, dP_Z(u)},$$

where the constant $B_Q$ satisfies

$$B_Q^2 = \sup_{z \in \text{supp} P_Z} B_{Q,z}^2. \quad (20)$$
Sensitivity: **Dataset-independent Bounds**

**Theorem**

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

$$\int \left| R_z \left( P_{\Theta}^{(Q,\lambda)} \right) - R_z \left( P_{\Theta|Z=z}^{(Q,\lambda)} \right) \right| dP_Z(z) \leq \sqrt{2B_Q^2 L(Z; \Theta)},$$

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

$$P_{\Theta}^{(Q,\lambda)}(A) = \int P_{\Theta|Z=z}^{(Q,\lambda)}(A) dP_Z(z) \quad (20)$$

and the constant $B_Q$ satisfies

$$B_Q^2 = \sup_{z \in \text{supp} P_Z} B_{Q,z}^2. \quad (21)$$
Universality

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

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathcal{B}(\mathcal{M}))} \int \mathcal{L}_z(\theta) dP(\theta),$$

subject to: $D \left( P \| P_{\Theta|Z=z}^{(Q,\lambda)} \right) \leq c$, with $c > 0$. (22b)

**Theorem**

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

$$\frac{dP_{\Theta|Z=z}^{(Q,\omega)}}{dQ}(\theta) = \exp \left( -K_{Q,z} \left( -\frac{1}{\omega} \right) - \frac{1}{\omega} \mathcal{L}_z(\theta) \right),$$

with $\omega \in (0, \lambda]$ such that $D \left( P_{\Theta|Z=z}^{(Q,\omega)} \| P_{\Theta|Z=z}^{(Q,\lambda)} \right) = c$. (23)
Table of Contents

1 The Problem of Supervised Learning

2 Empirical Risk Minimization with Relative Entropy Regularization

3 Probably Approximately Correct (PAC) Guarantees

4 Generalization Capabilities

5 Conclusions and Final Remarks
Generalization Capabilities: **Dataset Aggregation**

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

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

- **Constituent** datasets: $z_1$ and $z_2$

- **Aggregated** dataset: $z_0 = (z_1, z_2)$

- Interpretation 1: Training - Validation

- Interpretation 2: Decentralized learning

- Interpretation 3: Different data acquisition systems
Generalization Capabilities: **Dataset Aggregation**

- **Three datasets.** For all $i \in \{0, 1, 2\}$
  \[
  z_i = ( (x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \ldots, (x_{i,n_i}, y_{i,n_i}) ) \in (\mathcal{X} \times \mathcal{Y})^{n_i}, \text{ with } n_0 = n_1 + n_2
  \]  

- **Constituent** datasets: $z_1$ and $z_2$

- **Aggregated** dataset: $z_0 = (z_1, z_2)$

- Interpretation 1: Training - Validation

- Interpretation 2: Decentralized learning

- Interpretation 3: Different data acquisition systems

- Each dataset induces a different ERM-RER problem:
  \[
  \min_{P \in \Delta_{Q_i}(\mathcal{M}, \mathcal{B}(\mathcal{M}))} \mathcal{R}_{z_i}(P) + \lambda_i D(P\|Q_i).
  \]  

S. M. Perlaza (Inria, Sophia Antipolis)

Relative Entropy in Supervised Machine Learning

Generalization Capabilities: **Cross-Validation**

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

\[
Z_i = ((x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \ldots, (x_{i,n_i}, y_{i,n_i})) \in (X \times Y)^{n_i}, \text{ with } n_0 = n_1 + n_2 \quad (26)
\]

---

**Theorem (Cross-Validation Error)**

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

\[
R_{Z_i} \left( P_i(Q_j, \lambda_j) \right) - R_{Z_i} \left( P_i(Q_i, \lambda_i) \right)
\]

\[
= \lambda_i \left( D \left( P_i(Q_j, \lambda_j) \| P_i(Q_i, \lambda_i) \right) + D \left( P_i(Q_i, \lambda_i) \| Q_i \right) - D \left( P_i(Q_j, \lambda_j) \| Q_i \right) \right) . \quad (27)
\]
Generalization Capabilities: Quasi-Generalization Error

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

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

**Theorem (Quasi-Generalization Error)**

*For all* \( i \in \{0, 1, 2\} *, *

\[
R_{z_0} \left( P_{\Theta|Z=z_i}^{(Q_i,\lambda_i)} \right) - R_{z_0} \left( P_{\Theta|Z=z_0}^{(Q_0,\lambda_0)} \right) \\
= \lambda_0 \left( D(P_{\Theta|Z=z_i}^{(Q_i,\lambda_i)} \parallel P_{\Theta|Z=z_0}^{(Q_0,\lambda_0)}) + D(P_{\Theta|Z=z_0}^{(Q_0,\lambda_0)} \parallel Q_0) - D(P_{\Theta|Z=z_i}^{(Q_i,\lambda_i)} \parallel Q_0) \right) . \quad (29)
\]
Generalization Capabilities:
Homogeneous Priors and Proportional Regularization

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

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

- For all $i \in \{0, 1, 2\}$ and for all $\mathcal{A} \in \mathcal{B}(\mathcal{M})$, $Q(\mathcal{A}) = Q_i(\mathcal{A})$ (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 $\theta \in \mathcal{M}$, $L_{z_0}(\theta) = \frac{n_1}{n_0} L_{z_1}(\theta) + \frac{n_2}{n_0} L_{z_2}(\theta)$.***

*For all $\sigma$-finite measures $P \in \Delta(M, \mathcal{B}(M))$, $R_{z_0}(P) = \frac{n_1}{n_0} R_{z_1}(P) + \frac{n_2}{n_0} R_{z_2}(P)$.*
Generalization Capabilities: Homogeneous Priors and Proportional Regularization

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

\[
z_i = ((x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \ldots, (x_{i,n_i}, y_{i,n_i})) \in (\mathcal{X} \times \mathcal{Y})^{n_i}, \text{ with } n_0 = n_1 + n_2 \tag{30}
\]

- For all \( i \in \{0, 1, 2\} \) and for all \( A \in \mathcal{B}(\mathcal{M}), Q(A) = Q_i(A) \) (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 \( i \in \{1, 2\} \),

\[
R_{z_0} \left( P_{\Theta|Z=z_1}^{(Q,\lambda_1)} \right) - R_{z_0} \left( P_{\Theta|Z=z_2}^{(Q,\lambda_2)} \right) =
\]

\[
\lambda_0 \left( D\left( P_{\Theta|Z=z_1}^{(Q,\lambda_1)} \| P_{\Theta|Z=z_2}^{(Q,\lambda_2)} \right) + 2D\left( P_{\Theta|Z=z_1}^{(Q,\lambda_1)} \| Q \right) - D\left( P_{\Theta|Z=z_2}^{(Q,\lambda_2)} \| P_{\Theta|Z=z_1}^{(Q,\lambda_1)} \right) - 2D\left( P_{\Theta|Z=z_2}^{(Q,\lambda_2)} \| Q \right) \right).
\]
# Table of Contents

1. The Problem of Supervised Learning
2. Empirical Risk Minimization with Relative Entropy Regularization
3. Probably Approximately Correct (PAC) Guarantees
4. Generalization Capabilities
5. Conclusions and Final Remarks
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?
Chaining mutual information and tightening generalization bounds.

Rates of convergence for minimum contrast estimators.
*Pattern Recognition and Machine Learning*.  

Optimization methods for large-scale machine learning.  

_PAC-Bayesian supervised classification: The thermodynamics of statistical learning_, volume 56.


Cauchy, A.-L. (1847).

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

Differentially private empirical risk minimization.


A primer on PAC-Bayesian learning.

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

Maximum entropy discrimination.


How to escape saddle points efficiently.

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


Particle swarm optimization.

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


Adam: A method for stochastic optimization.


The sample average approximation method for stochastic discrete optimization.


*Concentration inequalities and model selection.*


A stochastic approximation method.


The calculus of M-estimation.


Information-theoretic analysis of generalization capability of learning algorithms.

Information-theoretic upper and lower bounds for statistical estimation.