Tutorial 2

Characterizing the Generalization Error of Machine Learning Algorithms via Information Measures

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

2024 IEEE Information Theory Workshop

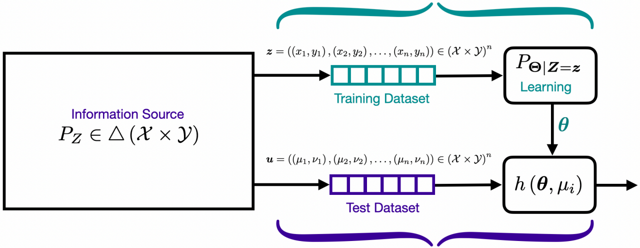
The 23th of November, 2024
Shenzhen, China

Slides for Part IV

Table of Contents (Part IV)

Table of Contents

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



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

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

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

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

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

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

Problem 1: ERM with Relative Entropy Regularization

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

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

Empirical Risk **Minimization** with Relative Entropy Regularization

The Gibbs Algorithm

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

Problem 1: ERM with Relative Entropy Regularization

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

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

Problem $1a$**:** ERM within a Neighborhood

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

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

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

Problem 2: Loss Maximization with Relative Entropy Regularization

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

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

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

Problem 2: Loss Maximization with Relative Entropy Regularization

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

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

Problem 2: Loss Maximization within a Neighbourhood

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

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

Problem 1:

Empirical Risk **Minimization** with Relative Entropy Regularization

Empirical Risk **Minimization** with Relative Entropy Regularization

Problem 1: ERM with Relative Entropy Regularization

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

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

Notation:

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

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

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

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

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

Empirical Risk **Minimization** with Relative Entropy Regularization

Problem 1: ERM with Relative Entropy Regularization

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

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

Notation:

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

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

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

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

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

Empirical Risk **Minimization** with Relative Entropy Regularization

Problem 1: ERM with Relative Entropy Regularization

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

Solution:

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

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

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

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

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

Empirical Risk **Minimization** with Relative Entropy Regularization

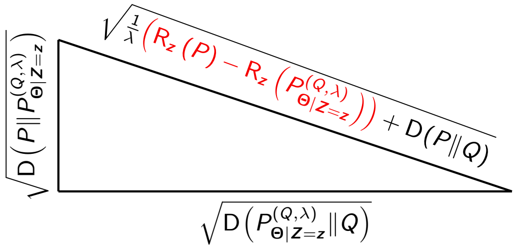


Figure: Geometric interpretation of the gap $\mathrm{R}_{z}\left(P\right)-\mathrm{R}_{z}\left(P^{(Q,\lambda)}_{\Theta|Z=z}\right)$.

Empirical Risk **Minimization** with Relative Entropy Regularization

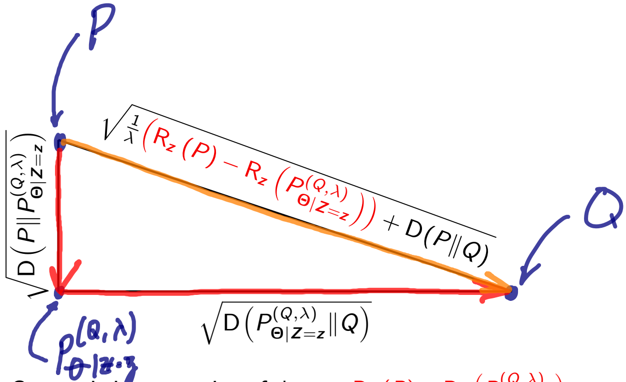


Figure: Geometric interpretation of the gap $R_z(P) - R_z\left(P^{(Q,\lambda)}_{\Theta|Z=z}\right)$.

Empirical Risk **Minimization** with Relative Entropy Regularization

Problem 1: ERM with Relative Entropy Regularization

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

Solution:

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

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

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

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

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

Problem 2:

Loss **Maximization** with Relative Entropy Regularization

Loss **Maximization** with Relative Entropy Regularization

Problem 2: Loss Maximization with Relative Entropy Regularization

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

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

Notation:

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

Theorem (Theorem 1 in [Zou-2024])

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

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

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

Loss **Maximization** with Relative Entropy Regularization

Problem 2: Loss Maximization with Relative Entropy Regularization

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

Solution:

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

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

Lemma (Theorem 6 in [Zou-2024])

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

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

Loss **Maximization** with Relative Entropy Regularization

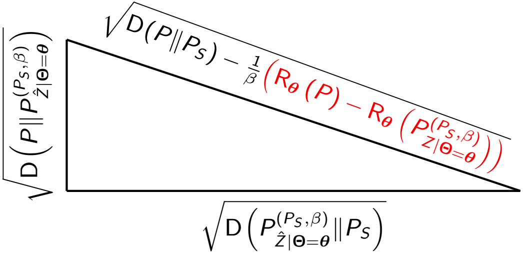


Figure: Geometric interpretation of the gap $\mathrm{R}_{\boldsymbol{\theta}}(P) - \mathrm{R}_{\boldsymbol{\theta}}\left(P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$.

Loss **Maximization** with Relative Entropy Regularization

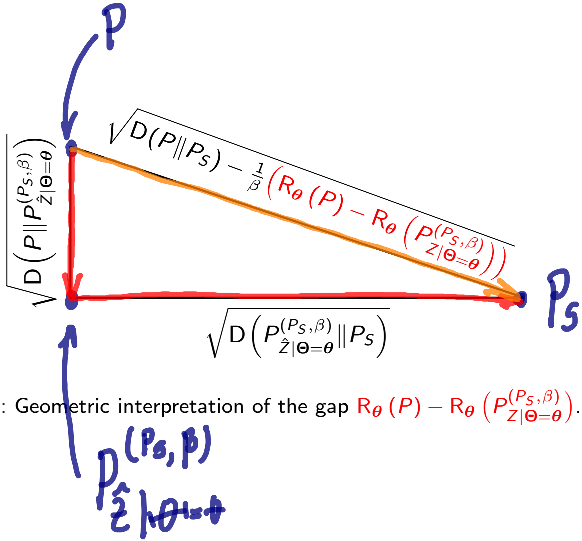


Figure: Geometric interpretation of the gap $R_{\theta}(P) - R_{\theta}\left(P^{(P_S,\beta)}_{Z|\Theta=\theta}\right)$.

Loss **Maximization** with Relative Entropy Regularization

Problem 2: Loss Maximization with Relative Entropy Regularization

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

Solution:

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

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

Theorem (Theorem 8 in [Zou-2024])

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

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

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

So far...

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

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

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

Theorem (Theorem 8 in [Zou-2024])

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

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

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

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

Table of Contents

The Method of Gaps

Definition (Expected Empirical Risk)

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

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

The Method of Gaps

Definition (Expected Empirical Risk)

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

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

Two **essential** observations:

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

The Method of Gaps

Definition (Expected Empirical Risk)

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

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

Two **essential** observations:

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

The Method of Gaps

Definition (Expected Empirical Risk)

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

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

Two **essential** observations:

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

Two-step Method:

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

The Method of Gaps

Definition (Expected Empirical Risk)

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

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

Two **essential** observations:

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

Two-step Method:

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

Definition (Expected Empirical Risk)

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

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

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

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

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

and

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

The Method of Gaps

Two variants:

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

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

The Method of **Algorithm-driven** Gaps

Generalization Error

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

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

The Method of **Algorithm-driven** Gaps

Generalization Error

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

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

Step 1:

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

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

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

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

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

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

Generalization Error

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

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

Step 2:

The Method of **Algorithm-driven** Gaps

Generalization Error

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

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

Step 2:

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

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

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

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

Generalization Error

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

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

Step 1:

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

The Method of **Data-driven** Gaps

Generalization Error

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

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

Step 1:

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

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

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

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

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

The Method of **Data-driven** Gaps

Generalization Error

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

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

Step 2:

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

Generalization Error

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

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

Step 2:

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

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

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

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

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

So far...

Generalization Error

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

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

So far...

Generalization Error

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

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

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

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

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

So far...

Generalization Error

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

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

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

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

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

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

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

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

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

Table of Contents

Expressions Obtained Via the Method of Gaps

- Particular **choices of the parameters**

Expressions Obtained Via the Method of Gaps

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

Expressions Obtained Via the Method of Gaps

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

Expressions Obtained Via the Method of Gaps

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

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

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

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

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

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

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

Statistical Hypothesis Test

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

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

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

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

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

Statistical Hypothesis Test

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

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

Expressions Obtained Via the Method of Gaps

Connections to Hypothesis Testing

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

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

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

Statistical Hypothesis Test

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

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

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

Expressions Obtained Via the Method of Gaps

Connections to Hypothesis Testing

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

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

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

Statistical Hypothesis Test

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

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

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

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

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

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

Statistical Hypothesis Test

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

Generalization Error

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Expressions Obtained Via the Method of Gaps

Connections to Information Measures

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

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

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

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

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

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

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

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

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

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

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

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

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

What if...

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

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

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

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

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

Generalization Error of the **Gibbs Algorithm**:

Corollary (Theorem 1 in [Aminian-2021])

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

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

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

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

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

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

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

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

What if...

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

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

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

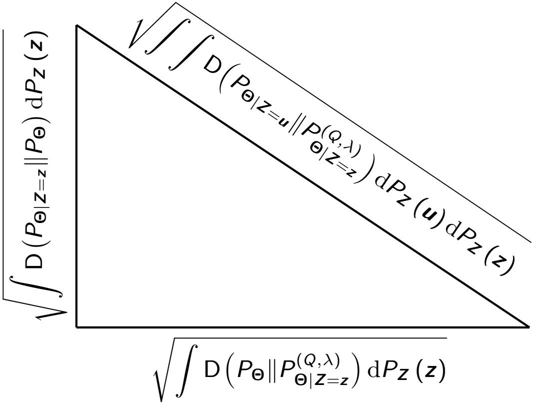
The generalization error $\overline{\overline{\mathsf{G}}}(P_{\Theta|Z}, P_{Z})$ *satisfies*

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

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

Expressions Obtained Via the Method of Gaps

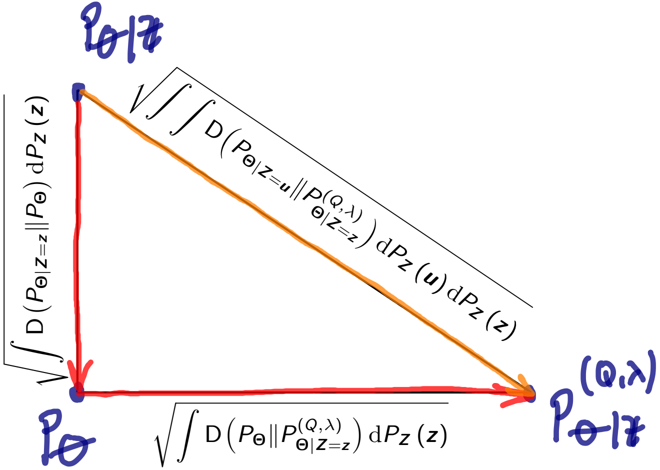
Connections to Euclidian Geometry



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

Expressions Obtained Via the Method of Gaps

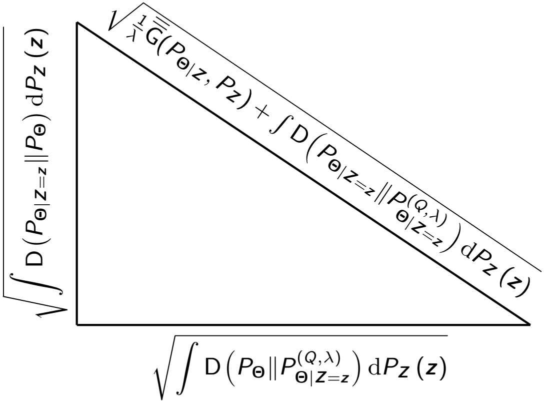
Connections to Euclidian Geometry



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

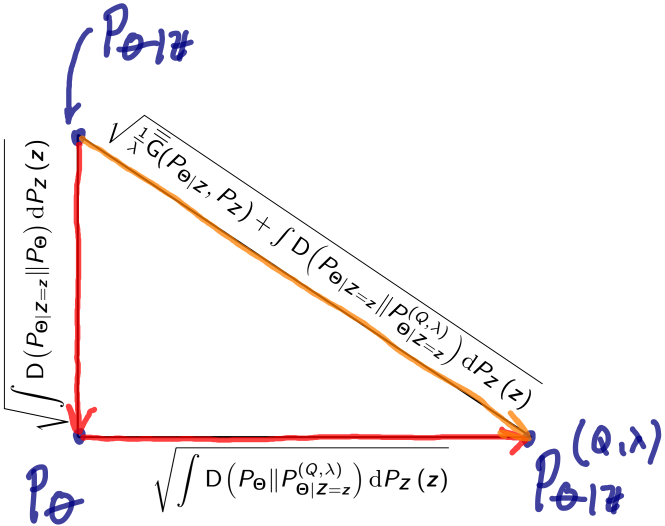
Expressions Obtained Via the Method of Gaps

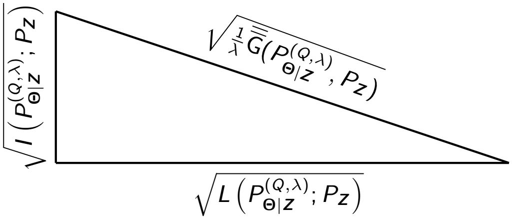
Connections to Euclidian Geometry



Expressions Obtained Via the Method of Gaps

Connections to Euclidian Geometry





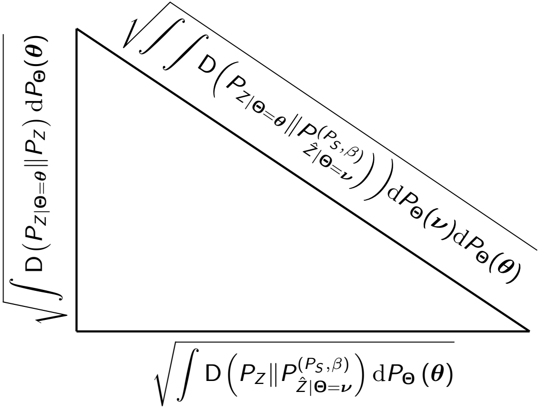
$\sqrt{I\left(P^{(Q,\lambda)}_{\Theta|Z}; P_Z\right)}$
$\sqrt{\frac{1}{\lambda}\overline{\overline{G}}(P^{(Q,\lambda)}_{\Theta|Z}, P_Z)}$
$\sqrt{L\left(P^{(Q,\lambda)}_{\Theta|Z}; P_Z\right)}$

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

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

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

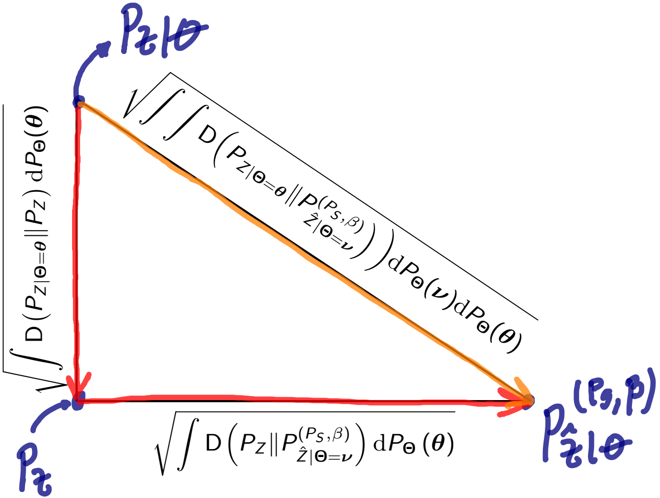
[Perlaza-2024b] Samir M. Perlaza and Xinying Zou. "**The Generalization Error of Machine Learning Algorithms**". November, 2024.



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

Expressions Obtained Via the Method of Gaps

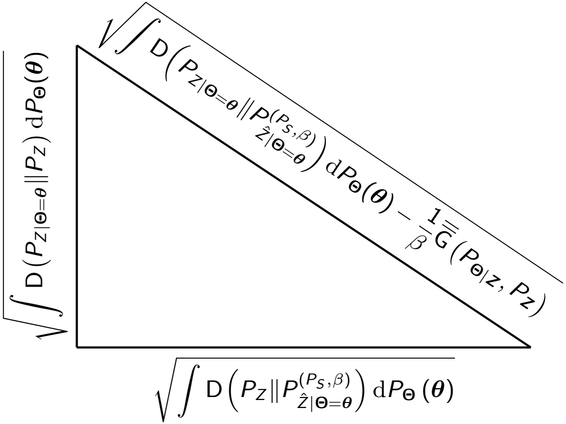
Connections to Euclidian Geometry



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

Expressions Obtained Via the Method of Gaps

Connections to Euclidian Geometry



Expressions Obtained Via the Method of Gaps

Connections to Euclidian Geometry

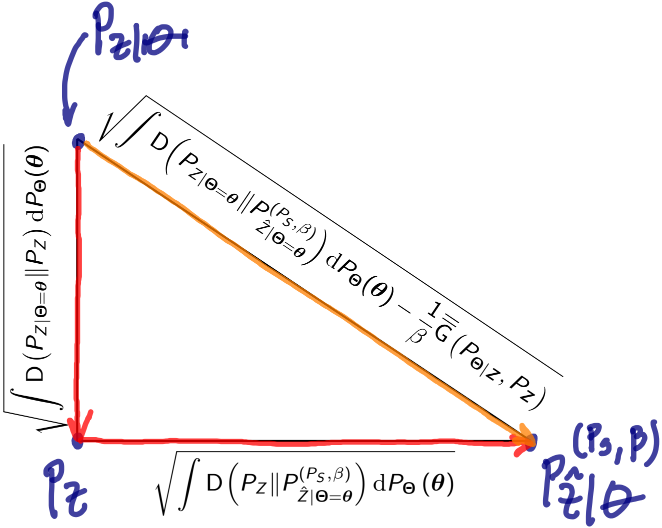


Table of Contents

Some Concluding Remarks

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

Some Concluding Remarks

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

Some Concluding Remarks

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

Examples

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

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

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

Examples

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

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

Thank you for your attention!

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

- This work appears in:

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