Information-Theoretic Limits for Inference, Learning, and Optimization

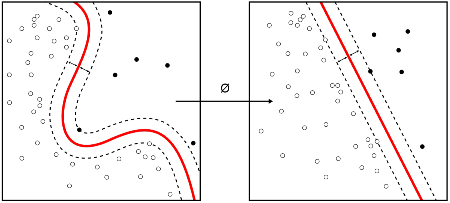
Part 3: Adaptive Data Analysis and Generalization Error

Jonathan Scarlett

Croucher Summer Course in Information Theory (CSCIT)
[July 2019]

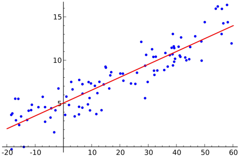
Typical Statistical Learning Goals

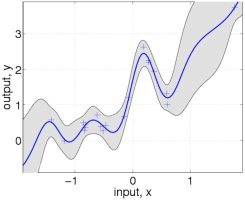
- **Classification:**



  - Spam detection, image classification, medical diagnosis, etc.

- **Regression:**

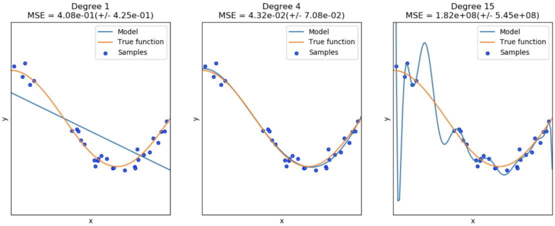




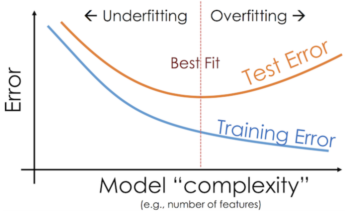
  - Stock price prediction, environmental monitoring, parameter optimization, etc.

Underfitting and Overfitting

- Example from [scikit-learn.com]:



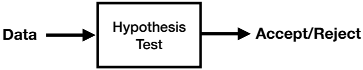
- Typical behavior of training/test error (at least classically) [ds100.org]:



- **Generalization error:** Difference between (average) test error and training error

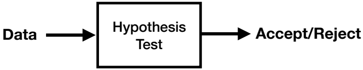
Hypothesis Testing and Adaptive Data Analysis

- **Scientific hypothesis testing:**

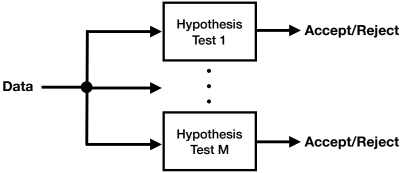


Hypothesis Testing and Adaptive Data Analysis

- **Scientific hypothesis testing:**

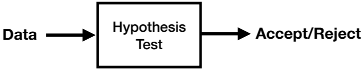


- **Scientific hypothesis testing of several hypotheses:**

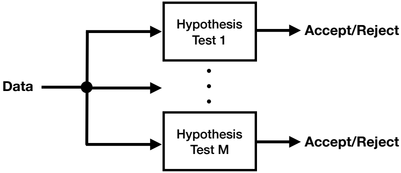


Hypothesis Testing and Adaptive Data Analysis

- **Scientific hypothesis testing:**



- **Scientific hypothesis testing of several hypotheses:**



- **Scientific adaptive data analysis:**

- **This talk:** Information-theoretic study of generalization error and spurious findings

(Very) Brief Overview of Some Classical Learning Theory

Statistical Learning

- **Basic notions:**
  - Input (feature) space $\mathcal{X}$
  - Output (label) space $\mathcal{Y}$
  - Function class $\mathcal{F}$ (e.g., set of all linear functions from $\mathcal{X}$ to $\mathcal{Y}$)
  - Loss function $\ell_f(\mathsf{x}, y)$ (e.g., squared loss $(y - f(\mathsf{x}))^2$)
  - Data set $\mathcal{D} = \{(\mathsf{x}_i, y_i)\}_{i=1}^n$ (i.i.d. from unknown $P_{\mathsf{X}Y}$)

Statistical Learning

- **Basic notions:**
  - Input (feature) space $\mathcal{X}$
  - Output (label) space $\mathcal{Y}$
  - Function class $\mathcal{F}$ (e.g., set of all linear functions from $\mathcal{X}$ to $\mathcal{Y}$)
  - Loss function $\ell_f(\mathsf{x}, y)$ (e.g., squared loss $(y - f(\mathsf{x}))^2$)
  - Data set $\mathcal{D} = \{(\mathsf{x}_i, y_i)\}_{i=1}^n$ (i.i.d. from unknown $P_{\mathsf{X}Y}$)
- **Measures of error:**
  - True average loss (true risk):

$$L(f) = \mathbb{E}[\ell_f(\mathsf{X}, Y)]$$

  - Empirical average loss (empirical risk):

$$L_{\mathcal{D}}(f) = \frac{1}{n}\sum_{i=1}^n \ell_f(\mathsf{x}_i, y_i)$$

  - A useful decomposition:

$$\underbrace{L(f)}_{\text{test error}} = \underbrace{L_{\mathcal{D}}(f)}_{\text{training error}} + \underbrace{\big(L(f) - L_{\mathcal{D}}(f)\big)}_{\text{generalization error}}.$$

Classical Generalization Bounds

- **PAC guarantee for bounded ℓ and finite $\mathcal{F}$:** If $n \geq \frac{2}{\epsilon^2} \log \frac{2|\mathcal{F}|}{\delta}$ then

$$L(F_{\rm erm}(\mathcal{D})) \leq \min_{f \in \mathcal{F}} L(f) + \epsilon$$

with probability at least $1 - \delta$.

  - Empirical risk minimization: $F_{\rm erm}(\mathcal{D}) = \arg\min_{f \in \mathcal{F}} L_{\mathcal{D}}(f)$
  - Analysis: Show that the true risk and empirical risk are close for **every** function in the class (uniform convergence)

Classical Generalization Bounds

- **PAC guarantee for bounded ℓ and finite $\mathcal{F}$:** If $n \geq \frac{2}{\epsilon^2} \log \frac{2|\mathcal{F}|}{\delta}$ then

$$L(F_{\rm erm}(\mathcal{D})) \leq \min_{f \in \mathcal{F}} L(f) + \epsilon$$

with probability at least $1 - \delta$.

  - Empirical risk minimization: $F_{\rm erm}(\mathcal{D}) = \arg\min_{f \in \mathcal{F}} L_{\mathcal{D}}(f)$
  - Analysis: Show that the true risk and empirical risk are close for **every** function in the class (uniform convergence)

- **PAC guarantee for 0-1 loss and infinite $\mathcal{F}$:** Similar to the case of finite classes, but with $n \geq C \cdot \frac{d_{\rm VC} + \log \frac{1}{\delta}}{\epsilon^2}$ ($d_{\rm VC}$ = VC dimension)

- **Other useful notions.** Rademacher complexity, algorithmic stability, PAC-Bayes, etc.

Proof Outline (Finite Function Class)

- **Concentration bound.** For any fixed $f \in \mathcal{F}$, we have

$$\mathbb{P}[|L_{\mathcal{D}}(f) - L(f)| \geq \epsilon_0] \leq 2e^{-2n\epsilon_0^2}$$

by Hoeffding's inequality

Proof Outline (Finite Function Class)

- **Concentration bound.** For any fixed $f \in \mathcal{F}$, we have

$$\mathbb{P}[|L_{\mathcal{D}}(f) - L(f)| \geq \epsilon_0] \leq 2e^{-2n\epsilon_0^2}$$

by Hoeffding's inequality

- **Union bound.** Applying the union bound gives

$$\mathbb{P}\left[\bigcup_{f \in \mathcal{F}} \{|L_{\mathcal{D}}(f) - L(f)| \geq \epsilon_0\}\right] \leq 2|\mathcal{F}|e^{-2n\epsilon_0^2}$$

Proof Outline (Finite Function Class)

- **Concentration bound.** For any fixed $f \in \mathcal{F}$, we have

$$\mathbb{P}[|L_{\mathcal{D}}(f) - L(f)| \geq \epsilon_0] \leq 2e^{-2n\epsilon_0^2}$$

by Hoeffding's inequality

- **Union bound.** Applying the union bound gives

$$\mathbb{P}\left[\bigcup_{f \in \mathcal{F}} \left\{|L_{\mathcal{D}}(f) - L(f)| \geq \epsilon_0\right\}\right] \leq 2|\mathcal{F}|e^{-2n\epsilon_0^2}$$

- **Re-arranging.** The above bound is at most δ provided that

$$n \geq \frac{1}{2\epsilon_0^2} \log \frac{2|\mathcal{F}|}{\delta}$$

Proof Outline (Finite Function Class)

- **Concentration bound.** For any fixed $f \in \mathcal{F}$, we have

$$\mathbb{P}[|L_{\mathcal{D}}(f) - L(f)| \geq \epsilon_0] \leq 2e^{-2n\epsilon_0^2}$$

by Hoeffding's inequality

- **Union bound.** Applying the union bound gives

$$\mathbb{P}\left[\bigcup_{f \in \mathcal{F}} \left\{|L_{\mathcal{D}}(f) - L(f)| \geq \epsilon_0\right\}\right] \leq 2|\mathcal{F}|e^{-2n\epsilon_0^2}$$

- **Re-arranging.** The above bound is at most δ provided that

$$n \geq \frac{1}{2\epsilon_0^2} \log \frac{2|\mathcal{F}|}{\delta}$$

- **Wrapping up.** Conditioned on the corresponding high probability event, we can easily show that $L(F_{\rm erm}) \leq L_{\min} + 2\epsilon_0$. Then set $\epsilon = 2\epsilon_0$.

Structural Risk Minimization

- **Multiple function classes.** If we have multiple function classes $\mathcal{F}_1, \dots, \mathcal{F}_M$, then the empirical risk minimizer for a "richer" $\mathcal{F}_m$ will tend to have lower training error, but higher generalization error

Structural Risk Minimization

- **Multiple function classes.** If we have multiple function classes $\mathcal{F}_1, \dots, \mathcal{F}_M$, then the empirical risk minimizer for a "richer" $\mathcal{F}_m$ will tend to have lower training error, but higher generalization error

- **Bayes optimal decision rule.**

$$F^* = \arg\min_{m=1,\dots,M} \arg\min_{f \in \mathcal{F}_m} L(f)$$

(can't be implemented if we don't know the true data distribution)

Structural Risk Minimization

- **Multiple function classes.** If we have multiple function classes $\mathcal{F}_1, \dots, \mathcal{F}_M$, then the empirical risk minimizer for a "richer" $\mathcal{F}_m$ will tend to have lower training error, but higher generalization error

- **Bayes optimal decision rule.**

$$F^* = \operatorname*{arg\,min}_{m=1,\dots,M} \operatorname*{arg\,min}_{f \in \mathcal{F}_m} L(f)$$

(can't be implemented if we don't know the true data distribution)

- **Structural risk minimization rule.**

$$F_{\mathrm{srm}} = \operatorname*{arg\,min}_{m=1,\dots,M} \operatorname*{arg\,min}_{f \in \mathcal{F}_m} L_n(f) + \overline{\mathrm{gen}}(\mathcal{F}_m)$$

where $\overline{\mathrm{gen}}(\mathcal{F}_m)$ is an upper bound on the generalization error for class $\mathcal{F}_m$

  - First term: Seek small training error
  - Second term: Regularization; penalize complex classes that may overfit

The Need for New Theoretical Tools

- **Key limitation of classical theory:** Overly pessimistic due to worst-case $P_{\mathsf{X}Y}$
  - Also often difficult to gain insight on specific learning algorithms and/or unbounded loss functions
  - (Note: More recent developments such as Rademacher complexity, PAC-Bayes, etc. are partially addressing some of these issues)

The Need for New Theoretical Tools

- **Key limitation of classical theory:** Overly pessimistic due to worst-case P_{XY}
  - Also often difficult to gain insight on specific learning algorithms and/or unbounded loss functions
  - (Note: More recent developments such as Rademacher complexity, PAC-Bayes, etc. are partially addressing some of these issues)

- **Modern challenges:**
  - Generalization performance can depend strongly on the data distribution
  - Would like theory to capture all ingredients of learning: Data distribution, function class, learning algorithm, and loss function
  - Many unsolved open problems (e.g., highly over-parametrized deep neural networks still generalize very well)
    - (Note: No claim of solving these using today's methods!)

Information Theory Approach

An Information-Theoretic Bound

- **Recap of notation:**
  - Data set $\mathcal{D} = \{(\mathsf{X}_i, Y_i)\}_{i=1}^n$, loss function $\ell_F(\mathsf{x}, y)$
  - True average loss $L(F)$, empirical average loss $L_{\mathcal{D}}(F)$
  - Data distribution $P_{\mathsf{X}Y}$, learning algorithm $P_{F|\mathcal{D}}$

- **Average generalization error:**

$$\begin{aligned}\mathrm{gen}(P_{\mathsf{X}Y}, P_{F|\mathcal{D}}) &= \mathbb{E}[L(F) - L_{\mathcal{D}}(F)] \\ &= \mathbb{E}\left[\ell_F(\mathsf{X}, Y) - \frac{1}{n}\sum_{i=1}^n \ell_F(\mathsf{X}_i, Y_i)\right]\end{aligned}$$

An Information-Theoretic Bound

- **Recap of notation:**
  - Data set $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, loss function $\ell_F(x, y)$
  - True average loss $L(F)$, empirical average loss $L_{\mathcal{D}}(F)$
  - Data distribution P_{XY}, learning algorithm $P_{F|\mathcal{D}}$

- **Average generalization error:**

$$\begin{aligned}\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) &= \mathbb{E}[L(F) - L_{\mathcal{D}}(F)] \\ &= \mathbb{E}\bigg[\ell_F(X, Y) - \frac{1}{n}\sum_{i=1}^n \ell_F(X_i, Y_i)\bigg]\end{aligned}$$

Claim. If $\ell_f(X, Y)$ is σ^2-subgaussian for all f, then [Russo and Zou, 2015]

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \le \sigma\sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

- σ^2-subgaussian: $\mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \le e^{\lambda^2\sigma^2/2}$ for all λ

Connection to Le Cam's Method

- Let $P_0(\mathsf{z})$ and $P_1(\mathsf{z})$ be two distributions on $\mathsf{z} = (\mathsf{x}, y)$

Connection to Le Cam's Method

- Let $P_0(\mathrm{z})$ and $P_1(\mathrm{z})$ be two distributions on $\mathrm{z} = (\mathrm{x}, y)$
- A very basic inequality (essentially by definition):

$$|\mathbb{P}_0[A] - \mathbb{P}_1[A]| \le \|P_0 - P_1\|_{\mathrm{TV}}$$

for any event A

  - Le Cam's Method: Use this inequality to lower bound hypothesis testing error probability in terms of TV norm; also extend to testing sets of distributions

Connection to Le Cam's Method

- Let $P_0(\mathrm{z})$ and $P_1(\mathrm{z})$ be two distributions on $\mathrm{z} = (\mathrm{x}, y)$
- A very basic inequality (essentially by definition):

$$|\mathbb{P}_0[A] - \mathbb{P}_1[A]| \le \|P_0 - P_1\|_{\mathrm{TV}}$$

for any event A

  - Le Cam's Method: Use this inequality to lower bound hypothesis testing error probability in terms of TV norm; also extend to testing sets of distributions

- Weakened version (via Pinsker's inequality):

$$|\mathbb{P}_0[A] - \mathbb{P}_1[A]| \le \sqrt{\frac{1}{2} D(P_1 \| P_0)}$$

(could also swap P_0 and P_1 on the right-hand side)

Connection to Le Cam's Method

- Let $P_0(\mathrm{z})$ and $P_1(\mathrm{z})$ be two distributions on $\mathrm{z} = (\mathrm{x}, y)$
- A very basic inequality (essentially by definition):

$$|\mathbb{P}_0[A] - \mathbb{P}_1[A]| \le \|P_0 - P_1\|_{\mathrm{TV}}$$

for any event A

- ▶ Le Cam's Method: Use this inequality to lower bound hypothesis testing error probability in terms of TV norm; also extend to testing sets of distributions

- Weakened version (via Pinsker's inequality):

$$|\mathbb{P}_0[A] - \mathbb{P}_1[A]| \le \sqrt{\frac{1}{2} D(P_1 \| P_0)}$$

(could also swap P_0 and P_1 on the right-hand side)

- Useful generalization: [Auer *et al.*, 1995]

$$|\mathbb{E}_0[a(\mathrm{z})] - \mathbb{E}_1[a(\mathrm{z})]| \le a_{\max} \sqrt{\frac{1}{2} D(P_1 \| P_0)}$$

for any function $a(\cdot)$ taking values in the range $[0, a_{\max}]$

Starts to look like what we want by setting (i) $P_1 \sim P_{\mathcal{D}F}$; (ii) $P_0 \sim P_{\mathcal{D}} \times P_F$; (iii) $a(\mathcal{D}, F) = L_{\mathcal{D}}(F)$. But not quite there (missing the crucial $\frac{1}{\sqrt{n}}$ dependence).

Proof Details

- Variational representation of relative entropy:

$$D(P\|Q) = \sup_{\tilde{g}} \Big(\mathbb{E}_P[\tilde{g}(Z)] - \log \mathbb{E}_Q[e^{\tilde{g}(Z)}] \Big)$$

Proof Details

- Variational representation of relative entropy:

$$D(P\|Q) = \sup_{\tilde{g}} \Big(\mathbb{E}_P[\tilde{g}(Z)] - \log \mathbb{E}_Q[e^{\tilde{g}(Z)}] \Big)$$

- Let $g_0(f, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^n \ell_f(X_i, Y_i)$, and $g(f, \mathcal{D}) = g_0(f, \mathcal{D}) - \mathbb{E}[g_0(f, \overline{\mathcal{D}})]$.
  - For fixed f and an independent data set $\overline{\mathcal{D}}$, $g(f, \overline{\mathcal{D}})$ is zero-mean and $\frac{\sigma^2}{n}$-subgaussian (by the i.i.d. data assumption), i.e., $\mathbb{E}[e^{\lambda g(f, \overline{\mathcal{D}})}] \le e^{\frac{\lambda^2 \sigma^2}{2n}}$
  - Hence, for $\overline{F}$ and $\overline{D}$ independent, $\mathbb{E}[e^{\lambda g(\overline{F}, \overline{\mathcal{D}})}] \le e^{\frac{\lambda^2 \sigma^2}{2n}}$
  - Also note that $\mathbb{E}[g(F, \mathcal{D})]$ is the generalization error

Proof Details

- Variational representation of relative entropy:

$$D(P\|Q) = \sup_{\tilde{g}} \Big(\mathbb{E}_P[\tilde{g}(Z)] - \log \mathbb{E}_Q[e^{\tilde{g}(Z)}] \Big)$$

- Let $g_0(f, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^n \ell_f(\mathsf{X}_i, Y_i)$, and $g(f, \mathcal{D}) = g_0(f, \mathcal{D}) - \mathbb{E}[g_0(f, \overline{\mathcal{D}})]$.
  - For fixed f and an independent data set $\overline{\mathcal{D}}$, $g(f, \overline{\mathcal{D}})$ is zero-mean and $\frac{\sigma^2}{n}$-subgaussian (by the i.i.d. data assumption), i.e., $\mathbb{E}[e^{\lambda g(f, \overline{\mathcal{D}})}] \le e^{\frac{\lambda^2 \sigma^2}{2n}}$
  - Hence, for $\overline{F}$ and $\overline{D}$ independent, $\mathbb{E}[e^{\lambda g(\overline{F}, \overline{\mathcal{D}})}] \le e^{\frac{\lambda^2 \sigma^2}{2n}}$
  - Also note that $\mathbb{E}[g(F, \mathcal{D})]$ is the generalization error

- Applying the above to $I(\mathcal{D}; F) = D(P_{\mathcal{D}F} \| P_{\mathcal{D}} \times P_F)$ and using $\tilde{g} = \lambda g$ gives

$$\begin{aligned} I(\mathcal{D}; F) &\ge \lambda \mathbb{E}[g(F, \mathcal{D})] - \log \mathbb{E}[e^{\lambda g(\overline{F}, \overline{\mathcal{D}})}] \\ &\ge \lambda \mathbb{E}[g(F, \mathcal{D})]] - \frac{\lambda^2 \sigma^2}{2n}. \end{aligned}$$

Proof Details

- Variational representation of relative entropy:

$$D(P\|Q) = \sup_{\tilde{g}} \left(\mathbb{E}_P[\tilde{g}(Z)] - \log \mathbb{E}_Q[e^{\tilde{g}(Z)}] \right)$$

- Let $g_0(f, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^n \ell_f(\mathsf{X}_i, Y_i)$, and $g(f, \mathcal{D}) = g_0(f, \mathcal{D}) - \mathbb{E}[g_0(f, \overline{\mathcal{D}})]$.
  - For fixed f and an independent data set $\overline{\mathcal{D}}$, $g(f, \overline{\mathcal{D}})$ is zero-mean and $\frac{\sigma^2}{n}$-subgaussian (by the i.i.d. data assumption), i.e., $\mathbb{E}[e^{\lambda g(f, \overline{\mathcal{D}})}] \le e^{\frac{\lambda^2 \sigma^2}{2n}}$
  - Hence, for $\overline{F}$ and $\overline{D}$ independent, $\mathbb{E}[e^{\lambda g(\overline{F}, \overline{\mathcal{D}})}] \le e^{\frac{\lambda^2 \sigma^2}{2n}}$
  - Also note that $\mathbb{E}[g(F, \mathcal{D})]$ is the generalization error

- Applying the above to $I(\mathcal{D}; F) = D(P_{\mathcal{D}F} \| P_{\mathcal{D}} \times P_F)$ and using $\tilde{g} = \lambda g$ gives

$$\begin{aligned} I(\mathcal{D}; F) &\ge \lambda \mathbb{E}[g(F, \mathcal{D})] - \log \mathbb{E}[e^{\lambda g(\overline{F}, \overline{\mathcal{D}})}] \\ &\ge \lambda \mathbb{E}[g(F, \mathcal{D})]] - \frac{\lambda^2 \sigma^2}{2n}. \end{aligned}$$

- Setting $\lambda = \frac{n\mathbb{E}[g(F, \mathcal{D})]]}{\sigma^2}$ gives

$$I(\mathcal{D}; F) \ge \frac{n\mathbb{E}[g(F, \mathcal{D})^2]}{2\sigma^2}$$

Re-arranging and noting $\mathbb{E}[g(F, \mathcal{D})] = \mathrm{gen}(P_{\mathsf{X}Y}, P_{F|\mathcal{D}})$ gives the desired result.

Examples

Example 1: Finite Function Class

- **Generalization bound:**

$$\mathrm{gen}(P_{\mathsf{X}Y}, P_{F|\mathcal{D}}) \le \sigma \sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

Example 1: Finite Function Class

- **Generalization bound:**

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \le \sigma \sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

- **Simple weakened version for finite $\mathcal{F}$:** [Xu and Raginsky, 2017]

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \le \sigma \sqrt{\frac{2H(F)}{n}}$$

  - Further bounding $H(F) \le \log |\mathcal{F}|$ gives classical bound for finite $\mathcal{F}$
  - But some learning algorithms may have much lower entropy! For instance, if some function in the class is "clearly best" then it has a much higher chance of being selected, so $H(F)$ is small.
  - Can also show that the general bound recovers the VC dimension based bound

Example 2: Quantization of Continuous Function Class

- **Generalization bound:**

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \le \sigma \sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

- **Quantization of output function for infinite $\mathcal{F}$:** Take initial output and "round" it to F restricted in some finite set, then use $I(\mathcal{D}; F) \le H(F)$

Example 2: Quantization of Continuous Function Class

- **Generalization bound:**

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \le \sigma \sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

- **Quantization of output function for infinite $\mathcal{F}$:** Take initial output and "round" it to F restricted in some finite set, then use $I(\mathcal{D}; F) \le H(F)$

- **Example.** [Xu and Raginsky, 2017]
  - If F is parametrized by some $\theta \in \mathbb{R}^d$ with $\|\theta\| \le B$, then quantizing to some $\hat{\theta}$ with $\|\hat{\theta} - \theta\| \le \frac{1}{\sqrt{n}}$ gives

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \le \sigma \sqrt{\frac{2d}{n} \log(2B\sqrt{dn})}$$

since it takes at most $(2B\sqrt{dn})^d$ points to always guarantee $\|\hat{\theta} - \theta\| \le \frac{1}{\sqrt{n}}$

Example 3: Randomized Selection

- **Generalization bound:**

$$\mathrm{gen}(P_{\mathsf{X}Y}, P_{F|\mathcal{D}}) \le \sigma\sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

- **Randomized selection example:** List the top m "best" functions in $\mathcal{F}$ and then select one of those uniformly at random [Russo and Zou, 2015]

Example 3: Randomized Selection

- **Generalization bound:**

$$\mathrm{gen}(P_{\mathsf{X}Y}, P_{F|\mathcal{D}}) \le \sigma\sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

- **Randomized selection example:** List the top m "best" functions in $\mathcal{F}$ and then select one of those uniformly at random [Russo and Zou, 2015]

- **Analysis.** Write

$$\begin{aligned} I(\mathcal{D}; F) &= H(F) - H(F|\mathcal{D}) \\ &\le \log|\mathcal{F}| - \log m \\ &= \log\frac{|\mathcal{F}|}{m} \end{aligned}$$

and hence

$$\mathrm{gen}(P_{\mathsf{X}Y}, P_{F|\mathcal{D}}) \le \sigma\sqrt{\frac{2}{n} \cdot \frac{\log|\mathcal{F}|}{m}}.$$

  - Multiplication of $\frac{1}{\sqrt{m}}$ compared to the standard finite-$\mathcal{F}$ bound

Example 3: Randomized Selection

- **Generalization bound:**

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \le \sigma \sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

- **Randomized selection example:** List the top m "best" functions in $\mathcal{F}$ and then select one of those uniformly at random [Russo and Zou, 2015]

- **Analysis.** Write

$$\begin{aligned} I(\mathcal{D}; F) &= H(F) - H(F|\mathcal{D}) \\ &\le \log|\mathcal{F}| - \log m \\ &= \log \frac{|\mathcal{F}|}{m} \end{aligned}$$

and hence

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \le \sigma \sqrt{\frac{2}{n} \cdot \frac{\log|\mathcal{F}|}{m}}.$$

  - Multiplication of $\frac{1}{\sqrt{m}}$ compared to the standard finite-$\mathcal{F}$ bound

- Increasing m tends to increase empirical risk but improve generalization

Example 4: Noisy ERM

- **Empirical risk minimization:**

$$\hat{f}_{\mathrm{erm}} = \arg\min_{f \in \mathcal{F}} L_{\mathcal{D}}(f),$$

i.e., choose the f with smallest training error

Example 4: Noisy ERM

- **Empirical risk minimization:**

$$\hat{f}_{\rm erm} = \arg\min_{f \in \mathcal{F}} L_{\mathcal{D}}(f),$$

i.e., choose the f with smallest training error

- **Noisy empirical risk minimization:**

$$\hat{f}_{\rm erm} = \arg\min_{f \in \mathcal{F}} L_{\mathcal{D}}(f) + N_f,$$

where N_f is an independent noise variable (e.g., exponential distribution)

- Can choose higher noise mean for more "a priori preferred" functions

Example 4: Noisy ERM

- **Empirical risk minimization:**

$$\hat{f}_{\mathrm{erm}} = \arg\min_{f \in \mathcal{F}} L_{\mathcal{D}}(f),$$

i.e., choose the f with smallest training error

- **Noisy empirical risk minimization:**

$$\hat{f}_{\mathrm{erm}} = \arg\min_{f \in \mathcal{F}} L_{\mathcal{D}}(f) + N_f,$$

where N_f is an independent noise variable (e.g., exponential distribution)

  - Can choose higher noise mean for more "a priori preferred" functions

- **Example risk guarantee:** For a countable function class $f_1, f_2, f_3, \dots$, under noisy ERM with $N_{f_i} \sim \mathrm{Exponential}(b_i)$, if $b_i = i^{1.1}/n^{1/3}$ then [Xu and Raginsky, 2017]

$$L(F) \le \min_i L(f_i) + \frac{\mathcal{I}^{1.1} + 3}{n^{1/3}}$$

where $\mathcal{I} = \arg\min_i L(f_i)$

Example 5: Iterative Algorithms

- **Generalization bound:**

$$\mathrm{gen}(P_{\mathsf{X}Y}, P_{F|\mathcal{D}}) \le \sigma\sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

Example 5: Iterative Algorithms

- **Generalization bound:**

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \le \sigma \sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

- **Iterative algorithms:** Suppose that we iteratively choose some F_j based on $\mathcal{D}$ and the outputs of previous stages $(F_1, \dots, F_{j-1})$
  - e.g., iterative optimization, going back to our data set because we didn't like what we obtained previously, etc.

Example 5: Iterative Algorithms

- **Generalization bound:**

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \le \sigma\sqrt{\frac{2}{n} I(\mathcal{D}; F)}$$

- **Iterative algorithms:** Suppose that we iteratively choose some F_j based on $\mathcal{D}$ and the outputs of previous stages $(F_1, \dots, F_{j-1})$
  - e.g., iterative optimization, going back to our data set because we didn't like what we obtained previously, etc.

- **Stage-wise upper bound on mutual information:**

$$\begin{aligned} I(\mathcal{D}; F_k) &\le I(\mathcal{D}; F_1, \dots, F_k) \\ &\le \sum_{j=1}^{k} I(\mathcal{D}; F_j | F_1, \dots, F_{j-1}) \end{aligned}$$

which resembles the upper bounding technique for channel coding with feedback

- See [Pensia *et al.*, 2018] for examples in stochastic gradient Langevin dynamics

Example 6: Gibbs Distribution

- **Mutual information regularization:** Since smaller $I(\mathcal{D}; F)$ reduces the generalization bound, one can consider using it as a regularizer:

$$\text{minimize}_{P_{F|\mathcal{D}}} \quad \mathbb{E}[L_{\mathcal{D}}(F)] + \frac{1}{\beta} I(\mathcal{D}; F)$$

Example 6: Gibbs Distribution

- **Mutual information regularization:** Since smaller $I(\mathcal{D}; F)$ reduces the generalization bound, one can consider using it as a regularizer:

$$\mathrm{minimize}_{P_{F|\mathcal{D}}} \ \ \mathbb{E}[L_{\mathcal{D}}(F)] + \frac{1}{\beta} I(\mathcal{D}; F)$$

- **A computable variant:** For fixed Q_F, upper bound $I(\mathcal{D}; F) \le D(P_{F|\mathcal{D}} \| Q_F | P_{\mathcal{D}})$; the resulting minimization

$$\mathrm{minimize}_{P_{F|\mathcal{D}}} \ \ \mathbb{E}[L_{\mathcal{D}}(F)] + \frac{1}{\beta} D(P_{F|\mathcal{D}} \| Q_F | P_{\mathcal{D}})$$

has a solution given by the Gibbs algorithm:

$$P_{F|\mathcal{D}}(f|D) = \frac{e^{-\beta L_D(f)}}{\mathbb{E}_Q[e^{-\beta L_D(F)}]}$$

Example 6: Gibbs Distribution

- **Mutual information regularization:** Since smaller $I(\mathcal{D}; F)$ reduces the generalization bound, one can consider using it as a regularizer:

$$\mathrm{minimize}_{P_{F|\mathcal{D}}} \;\; \mathbb{E}[L_{\mathcal{D}}(F)] + \frac{1}{\beta} I(\mathcal{D}; F)$$

- **A computable variant:** For fixed Q_F, upper bound $I(\mathcal{D}; F) \leq D(P_{F|\mathcal{D}} \| Q_F | P_{\mathcal{D}})$; the resulting minimization

$$\mathrm{minimize}_{P_{F|\mathcal{D}}} \;\; \mathbb{E}[L_{\mathcal{D}}(F)] + \frac{1}{\beta} D(P_{F|\mathcal{D}} \| Q_F | P_{\mathcal{D}})$$

has a solution given by the Gibbs algorithm:

$$P_{F|\mathcal{D}}(f|D) = \frac{e^{-\beta L_D(f)}}{\mathbb{E}_Q[e^{-\beta L_D(F)}]}$$

- **Generalization error:** [Xu and Raginsky, 2017]

$$\mathrm{gen}(P_{XY}, P_{F|\mathcal{D}}) \leq \frac{\beta}{2n}$$

Useful References

- **Original paper:** [Russo and Zou, 2015]

  https://arxiv.org/abs/1511.05219

- **Follow-up work:** [Xu and Raginsky, 2017]

  https://arxiv.org/abs/1705.07809

- (and several more – see "Cited By" on Google Scholar)