Differential Privacy and Generalization: Sharper Bounds, Theoretically Grounded Algorithms, and Thresholdout

Luca Oneto
www.lucaoneto.com
luca.oneto@unige.it

Summer School on Applied Harmonic Analysis,
Genoa, Italy, 24th July 2017

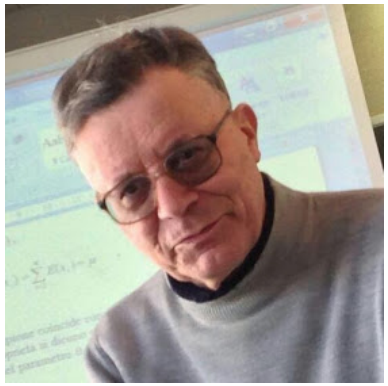
University of Genoa

Dibris
Polytechnic School and the School of Science

SmartLab
Research and Technology Transfer Laboratory

SmartLab

D. Anguita
Associate Prof.



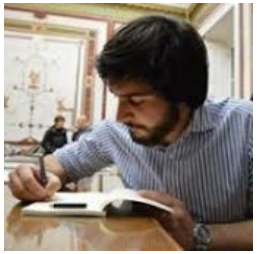
S. Ridella
Emeritus Prof.

L. Oneto
Assistant Professor

P. Sanetti
Res. Assistant

G. Clerico
Res. Assistant

A. Lulli
Postdoc



M. Cambiaso
Res. Assistant

I. Orlandi
PhD Stud.

E. Fumeo
PhD Stud.

F. Cipollini
PhD Stud.

Expertise

Scientific Research

- Topics
  - Neural Networks
  - Kernel Methods
  - Ensemble Methods
  - Statistical Learning Theory
  - Machine Learning
  - Data Mining
  - High Performance Computing
  - Big & Small Data Analysis
  - CBM, EDM, HAR, Sentiment Analysis, Cybercecurity
  - ...
- Publications
  - > 50 International High Ranked Journals
  - > 100 International High Ranked Conferences
  - ...

European Projects

- Basic Research
  - EC NeuroNet I & II - Network of Excellence on Neural Networks
  - EC RAIN - Redundant Array of Inexpensive Workstations for Neurocomputing
  - EC EUNITE - European Network of Excellence on Information Technology for Smart Adaptive Systems
  - EC-FET NiSIS - Nature-inspired Smart Information Systems
  - ...
- Applied Research
  - EC-H2020 IN2DREAMS
  - EC-H2020 In2Rail
  - EC-FP7 MAXBE
  - ...

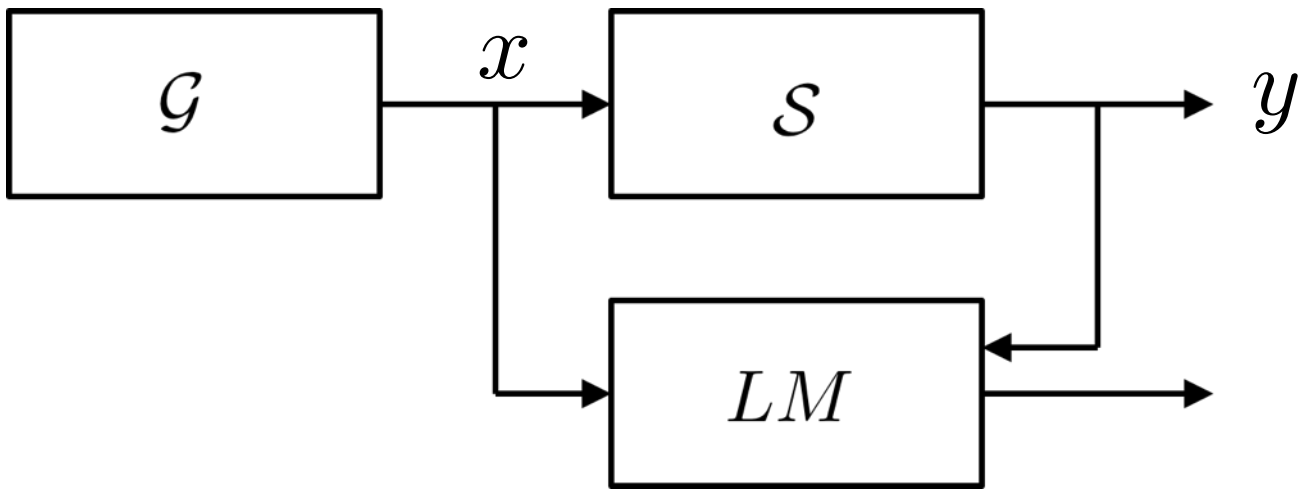
Technology Transfer

- aizoOn S.r.l.
- Ansaldo STS S.p.A.
- Brembo S.p.A
- Bombardier Transportation
- Cetena S.p.A.
- Damen Shipyards Group
- Ferrari S.p.A. - Scuderia Ferrari
- VarGroup
- ...

Privacy

- In the last years researchers have studied many ways to access data in a private way (aggregate, noise, etc.)
- Privacy is a bad thing from a data scientist point of view (we cannot access data if not aggregate, etc.)
- The breakthrough was to find a way to exploit privacy as a new regularization method and as a tool for better assessing the generalization performances of a learning algorithm

Supervised Learning



The only things available for learning is a set of examples of the mapping.

[1] Vapnik, V.N., 1998. Statistical learning theory. Wiley New York.

Deterministic Functions/Learning Algorithms

Given a set of data the algorithm always returns the same model.

$$f = \mathscr{A}(s)$$

The model is a function chosen in a set of functions: given the function and a point, the predicted output is always the same.

$$x \sim \mathfrak{P}_{\mathcal{X}}, \; y = f(x)$$

Randomized Functions

Given a set of data the algorithm always returns the same model.

$$\rho \leftarrow \mathscr{A}(s)$$

The model is a distribution over a set of functions. Given the model and a point, the predicted output may be different.

$$x \sim \mathfrak{P}_{\mathcal{X}}, \;\; f \sim \rho, \;\; y = f(x)$$

Randomized Learning Algorithms

Given a set of data the algorithm may return different models.

$$\boldsymbol{F} = \mathscr{A}(s)$$

The model can be a deterministic or randomized function. In our case the function is deterministic.

$$x \sim \mathfrak{P}_{\mathcal{X}}, \; y = \boldsymbol{F}(x)$$

Notation

$$x \in \mathcal{X}, \quad y \in \mathcal{Y}, \quad z \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}, \quad \mathfrak{P}_{\mathcal{X}}, \quad \mathfrak{P}_{\mathcal{Y}}, \quad \mathfrak{P}_{\mathcal{Z}}$$

$$\mathcal{Z}^n = \mathcal{S} \ni s = \{z_1, \cdots, z_n\} = \{(x_1, y_1), \cdots, (x_n, y_n)\} \quad \text{i.i.d.} \quad \mathfrak{P}_{\mathcal{S}}$$

$$\mathcal{Z} \ni \boldsymbol{Z} \sim \mathfrak{P}_{\mathcal{Z}}, \quad \mathcal{S} \ni \boldsymbol{S} \sim \mathfrak{P}_{\mathcal{S}}$$

$$\dot{s} = \{z_1, \cdots, z_{i-1}, \dot{z}_i, z_{i+1}, \cdots, z_n\} \quad \dot{z}_i \text{ i.i.d. } z_i$$

$$\breve{\mathcal{S}} \subseteq \mathcal{S}$$

$$f : \mathcal{X} \to \mathcal{Y}, \quad f \in \mathcal{F}$$

$$\breve{\mathcal{F}} \subseteq \mathcal{F}$$

$$\mathscr{A} : \mathcal{S} \to \mathcal{F}, \quad \mathfrak{P}_{\mathscr{A}}$$

$$\breve{D} : \mathcal{F} \to \breve{\mathcal{S}}$$

$$\ell : \mathcal{F} \times \mathcal{Z} \to [0, 1]$$

$$L(f) = \mathbb{E}_{\boldsymbol{Z}} \ell(f, \boldsymbol{Z}), \quad V(f) = \mathbb{E}_{\boldsymbol{Z}} [\ell(f, \boldsymbol{Z}) - L(f)]^2$$

$$\widehat{L}_n^s(f) = {}^1/_n \sum_{i=1}^{n} \ell(f, z_i), \quad \widehat{V}_n^s(f) = {}^1/_{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} [\ell(f, z_i) - \ell(f, z_j)]^2$$

[1] Oneto, Luca, Sandro Ridella, and Davide Anguita. "Differential privacy and generalization: Sharper bounds with applications." Pattern Recognition Letters 89 (2017): 31-38.
[2] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A., 2015b.Preserving statistical validity in adaptive data analysis, in: Annual ACM Symposium on Theory of Computing.

Goal

Estimate the true (generalization) error of the model based on the empirical data

$$\mathbb{P}\{|L(f) - \widehat{L}_n(f)| \geq \epsilon\} \leq \delta$$

$$|L(f) - \widehat{L}_n(f)| \leq \epsilon, \quad @(1-\delta)$$

$$\delta \leftrightarrow \epsilon$$

Differentially Private (DP) Randomized Learning Algorithms

A Randomized Learning Algorithm is (ϵ, δ)-DP if

$$\mathbb{P}_{\mathscr{A}}\left\{\mathscr{A}(s) \in \breve{\mathcal{F}}\right\} \leq e^{\epsilon}\mathbb{P}_{\mathscr{A}}\left\{\mathscr{A}(\dot{s}) \in \breve{\mathcal{F}}\right\} + \delta$$

$$\forall \breve{\mathcal{F}} \subseteq \mathcal{F},\ \forall s, \dot{s} \in \mathcal{S}$$

[1] Dwork, C., Roth, A., 2014. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science 9, 1–277.

Differentially Private (DP) Randomized Learning Algorithms

A Randomized Learning Algorithm is ϵ-DP if

$$\frac{\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(s)=f\}}{\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(\dot{s})=f\}} \leq e^{\epsilon},\ \forall f \in \mathcal{F},\ \forall s, \dot{s} \in \mathcal{S}$$

Proof: $\mathbb{P}_{\mathscr{A}}\left\{\mathscr{A}(s) \in \breve{\mathcal{F}}\right\} = \int_{\breve{\mathcal{F}}} \mathbb{P}_{\mathscr{A}}\{\mathscr{A}(s)=f\}df$

$$\leq \int_{\breve{\mathcal{F}}} e^{\epsilon}\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(\dot{s})=f\}df = e^{\epsilon}\mathbb{P}_{\mathscr{A}}\left\{\mathscr{A}(\dot{s}) \in \breve{\mathcal{F}}\right\}$$

[1] Oneto, Luca, Sandro Ridella, and Davide Anguita. "Differential privacy and generalization: Sharper bounds with applications." Pattern Recognition Letters 89 (2017): 31-38.

Hold Out, Compression, Complexity, Stability, and… Privacy (I)

The first option to estimate the generalization performance of an algorithm is to split the data

$$s = s_1 \cup s_2, \;\; s_1 \cap s_2 = \emptyset$$

$$\mathbb{P}_{\boldsymbol{S}_2}\left\{L(\mathscr{A}(s_1)) - \widehat{L}^{\boldsymbol{S}_2}_{|\boldsymbol{S}_2|}(\mathscr{A}(s_1)) \geq t\right\} \leq e^{-2|s_2|t^2}$$

[1] Hoeffding, Wassily. "Probability inequalities for sums of bounded random variables." Journal of the American statistical association 58.301 (1963): 13-30.
[2] Anguita D, Ghio A, Oneto L, Ridella S. In-sample and out-of-sample model selection and error estimation for support vector machines. IEEE Transactions on Neural Networks and Learning Systems. 2012 Sep;23(9):1390-406.

Hold Out, **Compression**, Complexity, Stability, and… Privacy (II)

Another option is to check how much the algorithm compresses the data

$s' \subseteq s$

$$\mathbb{P}_{\boldsymbol{S}}\left\{L(\mathscr{A}(\boldsymbol{S})) - \widehat{L}_n^{\boldsymbol{S}}(\mathscr{A}(\boldsymbol{S})) \geq t\right\} \leq n\binom{n}{|s'|}e^{-2nt^2}$$

[1] Floyd S, Warmuth M. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. Machine learning. 1995 Dec 1;21(3):269-304.
[2] Langford J, McAllester D. Computable shell decomposition bounds. Journal of Machine Learning Research. 2004;5(May):529-47.

Hold Out, Compression, **Complexity**, Stability, and… Privacy (III)

Another option is to check how large is the function space in which the algorithm chooses the solution

$$\mathbb{P}_{\boldsymbol{S}}\left\{L(\mathscr{A}(\boldsymbol{S})) - \widehat{L}_n^{\boldsymbol{S}}(\mathscr{A}(\boldsymbol{S})) \geq t\right\} \leq c_2 \mathtt{C}(\mathcal{F}) e^{-c_1 n t^2}$$

$$\mathbb{P}_{\boldsymbol{S}}\left\{L(\mathscr{A}(\boldsymbol{S})) - \widehat{L}_n^{\boldsymbol{S}}(\mathscr{A}(\boldsymbol{S})) \geq t\right\} \leq c_2 \mathtt{C}\left(\left\{f : f \in \mathcal{F}, \widehat{L}_n^{s}(f) \leq c_3\right\}\right) e^{-c_1 n t^2}$$

$c_3(L, \mathtt{C}, t, n)$

$$\mathtt{C}(\mathcal{F}) : \; n^{d_{VC}(\mathcal{F})}, \; e^{n \mathtt{R}^2(\mathcal{F})}$$

[1] V. N. Vapnik, Statistical learning theory, Wiley-Interscience, 1998.
[2] L. Oneto, A. Ghio, S. Ridella, D. Anguita, Global rademacher com- plexity bounds: From slow to fast convergence rates, Neural Processing Letters 43 (2) (2015) 567–602.
[3] L. Oneto, A. Ghio, S. Ridella, D. Anguita, Local rademacher complex- ity: Sharper risk bounds with and without unlabeled samples, Neural Networks 65 (2015) 115–125.

Hold Out, Compression, Complexity, **Stability**, and… Privacy (IV)

Another way is to check how close the functions chosen by the algorithm are

$$|\ell(\mathscr{A}(s),\cdot) - \ell(\mathscr{A}(\dot{s}),\cdot)|_\infty \leq \beta$$

$$\mathbb{P}_{\boldsymbol{S}}\left\{L(\mathscr{A}(\boldsymbol{S})) - \widehat{L}_n^{\boldsymbol{S}}(\mathscr{A}(\boldsymbol{S})) \geq t\right\} \leq c_2 e^{n\beta^2 - c_1 n t^2}$$

[1] O. Bousquet, A. Elisseeff, Stability and generalization, The Journal of Machine Learning Research 2 (2002) 499–526.
[2] L. Oneto, A. Ghio, S. Ridella, D. Anguita, Fully empirical and data- dependent stability-based bounds, IEEE Transactions on Cybernetics 45 (9) (2015) 1913–1926.
[3] Maurer A. A Second-order Look at Stability and Generalization. InConference on Learning Theory 2017 Jun 18 (pp. 1461-1475).

DP Main Result

$$\text{If } \mathbb{P}_{\boldsymbol{S}}\{\boldsymbol{S} \in \breve{D}(f)\} \leq \beta$$

$$\forall f \in \mathcal{F} \text{ and } \epsilon \leq \sqrt{\ln\left(1/\beta\right)/2n}$$

$$\rightarrow \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{\boldsymbol{S} \in \breve{D}(\boldsymbol{F})\} \leq 3\sqrt{\beta}$$

Proof: rather technical…

[1] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, A. Roth, Preserving statistical validity in adaptive data analysis, in: Symposium on Theory of Computing, 2015.

Hoeffding-type Bounds

$$\epsilon \leq \sqrt{t^2 - \ln(2)/2n}$$

$$\rightarrow \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{|L(\boldsymbol{F}) - \widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})| \geq t\} \leq 3\sqrt{2}e^{-nt^2}$$

$$O\left(1/\sqrt{n}\right)$$

Proof: $\mathbb{P}_{\boldsymbol{S}}\{L(f) - \widehat{L}_n^{\boldsymbol{S}}(f) \geq t\} \leq e^{-2nt^2}$

$$\breve{D}(f) = \{s \in \mathcal{S} : L(f) - \widehat{L}_n^{\boldsymbol{S}}(f) > t\}$$

$$\beta = e^{-2nt^2}$$

[1] Hoeffding, Wassily. Probability inequalities for sums of bounded random variables. Journal of the American statistical association 58.301 (1963): 13-30.
[2] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, A. Roth, Preserving statistical validity in adaptive data analysis, in: Symposium on Theory of Computing, 2015.
[3] L. Oneto, S. Ridella, D. Anguita, Differential privacy and generalization: Sharper bounds with applications, Pattern Recognition Letters 89 (2017) 31–38.

Chernoff and Bennett-type Bounds

$$\epsilon \leq \sqrt{t^2 - \ln(2)/2n}$$

$$\rightarrow \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{|L(\boldsymbol{F}) - \widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})| \geq \sqrt{6L(\boldsymbol{F})}t\} \leq 3\sqrt{2}e^{-nt^2}$$

$$O\left(1/\sqrt{n}\right) \div O\left(1/n\right)$$

[1] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, The Annals of Mathematical Statistics 23 (4) (1952) 493–507.
[2] A. Maurer, M. Pontil, Empirical bernstein bounds and sample variance penalization, in: Conference on Learning Theory, 2009.
[3] L. Oneto, S. Ridella, D. Anguita, Differential privacy and generalization: Sharper bounds with applications, Pattern Recognition Letters 89 (2017) 31–38.

Chernoff and **Bennett**-type Bounds

$$\epsilon \leq \sqrt{t^2 - {}^{\ln(3)}/_{2n}}$$

$$\rightarrow \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}} \left\{ \left| L(\boldsymbol{F}) - \widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) \right| \geq \sqrt{4\widehat{V}_n^{\boldsymbol{S}}(\boldsymbol{F})t} + \frac{14nt^2}{3(n-1)} \right\} \leq 3\sqrt{3}e^{-nt^2}$$

$$O\left({}^{1}/_{\sqrt{n}}\right) \div O\left({}^{1}/_{n}\right)$$

[1] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, The Annals of Mathematical Statistics 23 (4) (1952) 493–507.
[2] A. Maurer, M. Pontil, Empirical bernstein bounds and sample variance penalization, in: Conference on Learning Theory, 2009.
[3] L. Oneto, S. Ridella, D. Anguita, Differential privacy and generalization: Sharper bounds with applications, Pattern Recognition Letters 89 (2017) 31–38.

Clopper-Pearson (Binary Classification)

$$\epsilon \leq \sqrt{\ln\left(1/2\delta\right)/2n}$$

$$\rightarrow \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{\mathtt{Q}[\delta; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}), n - n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) + 1] \leq L(\boldsymbol{F})$$

$$\leq \mathtt{Q}[1-\delta; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) + 1, n - n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})]\} \leq 3\sqrt{2\delta}$$

$$O\left(1/\sqrt{n}\right) \div O\left(1/n\right)$$

[1] C. J. Clopper, E. S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial, Biometrika 26 (4) (1934) 404– 413.
[2] L. Oneto, S. Ridella, D. Anguita, Differential privacy and generalization: Sharper bounds with applications, Pattern Recognition Letters 89 (2017) 31–38.

Clopper-Pearson (Regression)

$$\epsilon \leq \sqrt{\ln\left(1/2\delta\right)/2n}$$

$$\rightarrow \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\left\{\mathtt{Q}\left[\delta; \sum_{i=1}^{n}\left[\ell(\boldsymbol{F},\boldsymbol{Z}_i) \geq u_i\right], n - \sum_{i=1}^{n}\left[\ell(\boldsymbol{F},\boldsymbol{Z}_i) \geq u_i\right] + 1\right] \leq L(\boldsymbol{F})\right.$$

$$\left.\leq \mathtt{Q}\left[1-\delta; \sum_{i=1}^{n}\left[\ell(\boldsymbol{F},\boldsymbol{Z}_i) \geq u_i\right] + 1, n - \sum_{i=1}^{n}\left[\ell(\boldsymbol{F},\boldsymbol{Z}_i) \geq u_i\right]\right]\right\} \leq 3\sqrt{2\delta}$$

Proof: $\begin{cases} \mathbb{P}\{u = \alpha, \alpha \in [0,1]\} = 1 \\ \mathbb{P}\{u = \alpha, \alpha \notin [0,1]\} = 1 \end{cases}$

$$\mathbb{P}\left\{h \geq u\right\} = \mathbb{E}_{h,u}\left\{h \geq u\right\}$$
$$= \mathbb{E}_h\left\{\mathbb{E}_u\left\{h \geq u\right\}\right\} = \mathbb{E}_h\left\{h\right\}$$

$$O\left(1/\sqrt{n}\right) \div O\left(1/n\right)$$

[1] L. Oneto, S. Ridella, D. Anguita, Differential privacy and generalization: Sharper bounds with applications, Pattern Recognition Letters 89 (2017) 31–38.
[2] X. Chen, A link between binomial parameters and means of bounded random variables, arXiv preprint arXiv:0802.3946.

Example: DP Random Forest (RF) (I)

www.openml.org

Abb.	ID	Name	n	d	Abb.	ID	Name	n	d
D01	40	sonar	208	61	D02	59	ionosphere	351	35
D03	785	wind_correlations	45	47	D04	882	pollution	60	16
D05	1104	leukemia	72	7130	D06	1446	CostaMadre1	296	38
D07	1453	PieChart3	1077	38	D08	1458	arcene	200	10001
D09	1485	madelon	2600	501	D10	1566	hill-valley	1212	101
D11	37	diabetes	768	9	D12	1005	glass	214	10
D13	1494	qsar-biodeg	1055	42	D14	1134	OVA_Kidney	1545	10937
D15	1217	Click_prediction_small	149639	12	D16	1149	AP_Ovary_Kidney	458	10937
D17	907	chscase_census4	400	8	D18	976	kdd_JapaneseVowels	9961	15
D19	1443	PizzaCutter1	61	38	D20	871	pollen	3848	6

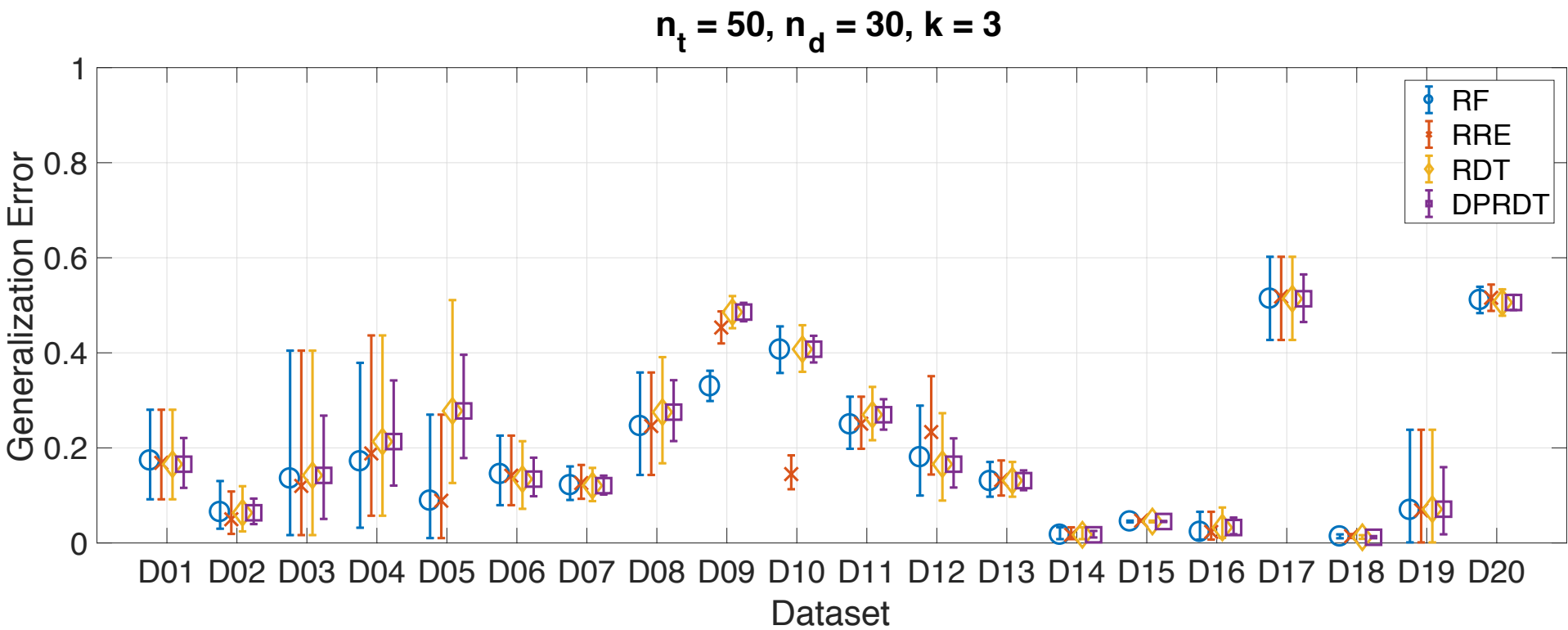
Example: DP RF (II)

- the Random Forests (RF): the original RF formulation
- the Random Rotation Ensembles (RRE): a recent improvement over the original RF
- the Random Decision Trees (RDT): a fully random RF implementation which is faster to be trained
- the Differentially Private RDT (DPRDT): a RDT formulation which is also DP

[1] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.
[2] R. Blaser, P. Fryzlewicz, Random rotation ensembles, Journal of Ma- chine Learning Research 17 (4) (2015) 1–15.
[3] M. Bojarski, A. Choromanska, K. Choromanski, Y. LeCun, Differentially-and non-differentially-private random decision trees, in: arXiv preprint arXiv:1410.6973, 2014.

Example: DP RF (III)

- kCV for RF, RRE, and RDT
- DP for DPRDT



Randomized Functions or Randomized Algorithms?

For studying Randomized Algorithms we have different options

- Hold out
- Stability
- DP

For studying Randomized Functions, instead we only have one powerful option

- PAC-Bayes theory

PAC-Bayes Theory (I)

$$\pi : \text{ Prior over } \mathcal{F}$$

$$\rho : \text{ Posterior over } \mathcal{F}$$

$$G_\rho(X) : x \sim \mathfrak{P}_\mathcal{X}, \; f \sim \rho, \; y = f(x)$$

$$B_\rho(X) : x \sim \mathfrak{P}_\mathcal{X}, y = \mathbb{E}_\rho\{f(X)\}$$

$$L(B\rho) \leq 2L(G\rho)$$

[1] Germain, P., et al. Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm. Journal of Machine Learning Research. 2015.
[2] Germain, Pascal, et al. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. The Journal of Machine Learning Research 16.1 (2015): 787-860.
[3] McAllester D. A. Some pac-bayesian theorems. Computational learning theory. 1998
Langford J. Tutorial on practical prediction theory for classification."Journal of machine learning research. 2005 42
[4]Germain P. Lacasse A. Laviolette F. Marchand M. PAC-Bayesian learning of linear classifiers. International Conference on Machine Learning. 2009

PAC-Bayes Theory (II)

$$\mathbb{P}\left\{ \mathtt{kl}\left[\widehat{L}_n^{\boldsymbol{S}}(G_Q)||L(G_Q)\right] \geq \frac{\mathtt{KL} + \ln\left[\frac{2\sqrt{n}}{\delta}\right]}{n} \right\} \leq \delta$$

$$\mathtt{KL}(\rho||\pi) = \mathbb{E}_\rho\left\{\ln\left(\frac{\rho}{\pi}\right)\right\}$$

$$O\left(\sqrt{\ln(n)/n}\right) \div O\left(\ln(n)/n\right)$$

[1] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79-86.
[2] Bégin, L., Germain, P., Laviolette, F., & Roy, J. F. (2016). PAC-Bayesian Bounds based on the Rényi Divergence. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (pp. 435-444).

Catoni's Result: Catoni Randomized Function (CRF)

$$\pi(f) = \frac{1}{Z}\exp(-\gamma L(f))$$

$$\rho(f) = \frac{1}{Z'}\exp(-\gamma \widehat{L}_n^s(f))$$

$$\mathtt{KL}(\rho||\pi) \leq \frac{\gamma^2}{2n} + \gamma\sqrt{\frac{2\ln\left[\frac{2\sqrt{n}}{\delta}\right]}{n}}, \quad \text{at: } (1-\delta)$$

[1] O.Catoni.Pac-bayesian supervised classification: The thermodynamics of statistical learning. arXiv preprint arXiv:0712. 0248, 2007.
[2] Lever G. Laviolette F. Shawe-Taylor J. Tighter PAC-Bayes bounds through distribution-dependent priors. Theoretical Computer Science. 2013
[3] Oneto, Luca, Davide Anguita, and Sandro Ridella. PAC-bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis. Pattern Recognition Letters 80 (2016): 200-207.

Why not a Catoni Randomized Algorithm (CRA)?

Instead of building CDF based on the Catoni's posterior we can think about a Randomized Algorithm which chooses, inside our space of function, the best function (the one with the smallest empirical error) perturbed with the Catoni's noise (CRA).

[1] L. Oneto, S. Ridella, D. Anguita, Differential privacy and generalization: Sharper bounds with applications, Pattern Recognition Letters 89 (2017) 31–38.

CRA is DP

$$\frac{\mathbb{P}\{\mathscr{A}(s)=f\}}{\mathbb{P}\{\mathscr{A}(\dot{s})=f\}}$$

$$= \frac{e^{-\frac{\gamma}{n}\sum_{i=1}^{n}\ell(f,z_i)}}{\sum_{f_1\in\mathcal{F}} e^{-\frac{\gamma}{n}\sum_{i=1}^{n}\ell(f_1,z_i)}} \frac{\sum_{f_1\in\mathcal{F}} e^{-\frac{\gamma}{n}\left(\sum_{i=1,i\neq j}^{n}\ell(f_1,z_i)+\ell(f_1,\dot{z}_j)\right)}}{e^{-\frac{\gamma}{n}\left(\sum_{i=1,i\neq j}^{n}\ell(f,z_i)+\ell(f,\dot{z}_j)\right)}}$$

$$\leq \frac{e^{0}}{\sum_{f_1\in\mathcal{F}} e^{-\frac{\gamma}{n}\sum_{i=1,i\neq j}^{n}\ell(f_1,z_i)} e^{-\frac{\gamma}{n}}} \frac{\sum_{f_1\in\mathcal{F}} e^{-\frac{\gamma}{n}\sum_{i=1,i\neq j}^{n}\ell(f_1,z_i)} e^{0}}{e^{-\frac{\gamma}{n}}}$$

$$= e^{\frac{2\gamma}{n}}.$$

[1] L. Oneto, S. Ridella, D. Anguita, Differential privacy and generalization: Sharper bounds with applications, Pattern Recognition Letters 89 (2017) 31–38.

CRF and CRA Generalization Properties

CRA

$$O\left(1/\sqrt{n}\right) \div O\left(1/n\right)$$

$$\gamma = 1/2\sqrt{n \ln\left(3\sqrt{2}/2\delta\right)}$$

$$\rightarrow \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\left\{\mathtt{Q}\left[\frac{\delta^2}{18}; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}), n - n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) + 1\right] \leq L(\boldsymbol{F})\right.$$

$$\left.\leq \mathtt{Q}\left[1 - \frac{\delta^2}{18}; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) + 1, n - n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})\right]\right\} \leq \delta$$

CRF

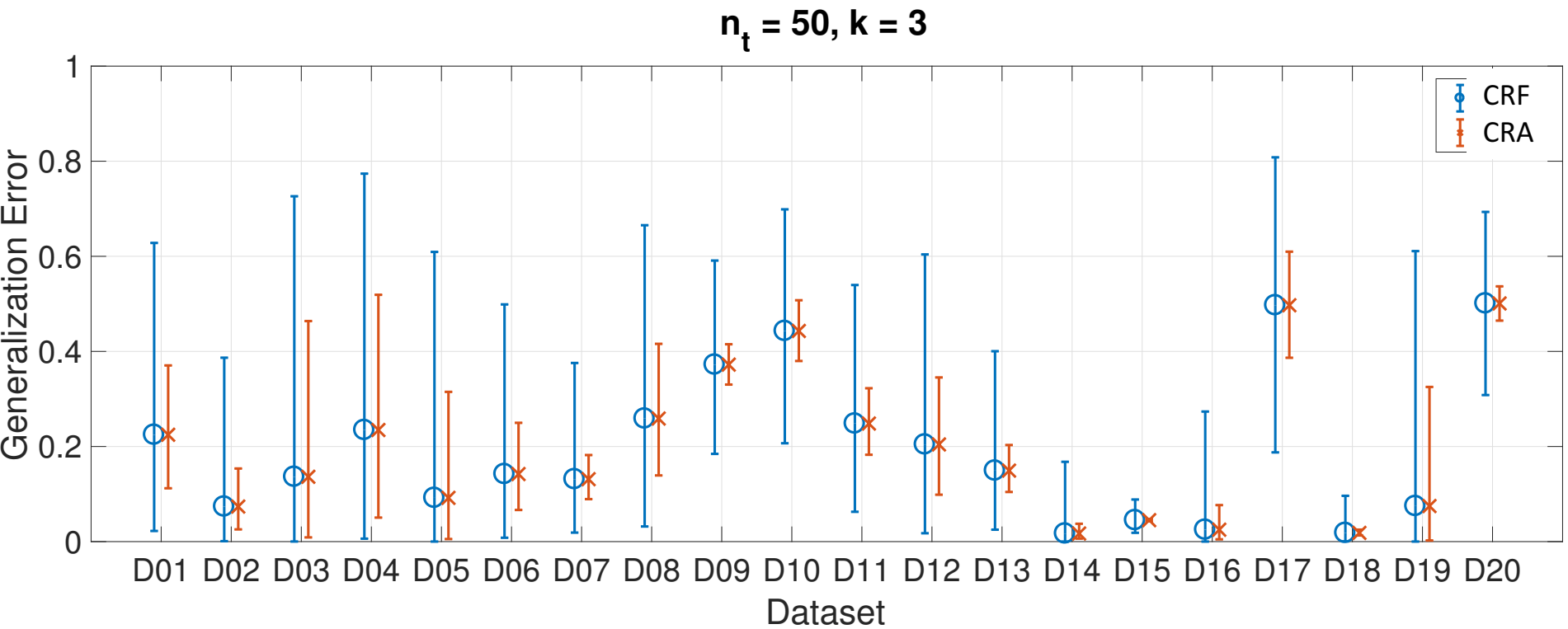
$$O\left(\sqrt[4]{\ln(n)/n}\right) \div O\left(\sqrt{\ln(n)/n}\right)$$

$$\gamma = 1/2\sqrt{n \ln\left(3\sqrt{2}/2\delta\right)}$$

$$\rightarrow \mathbb{P}_{\boldsymbol{S}}\left\{\mathtt{kl}[\widehat{L}_n^{\boldsymbol{S}}(G_Q)||L(G_Q)] \geq \frac{\ln\left(\frac{3\sqrt{2}}{\delta}\right)}{8n} + \sqrt{\frac{2\ln\left(\frac{3\sqrt{2}}{\delta}\right)\ln\left(\frac{2\sqrt{n}}{\delta}\right)}{4n}} + \frac{\ln\left(\frac{2\sqrt{n}}{\delta}\right)}{n}\right\} \leq 2\delta.$$

Example: CRF and CRA

Functions space counts of trees built with an hold out set.



Non-Adaptive Data Analysis (I)

NAS: the non-adaptive setting is the case when the procedures for building our models exploit just the training set.

Algorithm 1: Union Bound for the NAS

Input: s_t, s_h, and $\mathscr{P}_1, \cdots, \mathscr{P}_m$
Output: $\widehat{L}_n^{s_h}(f_1), \cdots, \widehat{L}_n^{s_h}(f_m)$
for $i \leftarrow 1$ **to** m **do**
  $f_i = \mathscr{P}_i(s_t)$ and compute $\widehat{L}_n^{s_h}(f_i)$;

[1] Oneto L, Ridella S, Anguita D. Differential privacy and generalization: Sharper bounds with applications. Pattern Recognition Letters. 2017 Apr 1;89:31-8.

Non-Adaptive Data Analysis (II)

NAS: the non-adaptive setting is the case when the procedures for building our models exploit just the training set.

Then we can use the Bonferroni Correction:

$$\mathbb{P}_{\boldsymbol{S}_h}\left\{\exists i \in \mathcal{I}_m : \left|L(\mathscr{P}_i(s_t)) - \widehat{L}_n^{\boldsymbol{S}_h}(\mathscr{P}_i(s_t))\right| \geq \sqrt{\frac{\ln\left(\frac{2m}{\delta}\right)}{2n}}\right\} \leq \delta$$

$$O\left(\sqrt{\ln(m)/n}\right) \div O\left(\ln(m)/n\right)$$

[1] Oneto L, Ridella S, Anguita D. Differential privacy and generalization: Sharper bounds with applications. Pattern Recognition Letters. 2017 Apr 1;89:31-8.

Adaptive Data Analysis (I)

AS: the adaptive setting is the case when the procedures for building our models exploit both the training set and the performance of the procedure at previous step over the hold out set.

Algorithm 1: Hold out for the AS

Input: s_t, s_h, and $\mathscr{P}_1, \cdots, \mathscr{P}_m$

Output: $\widehat{L}_n^{s_h^1}(f_1), \cdots, \widehat{L}_n^{s_h^m}(f_m)$

Split s_h in s_h^i with $i \in \mathcal{I}_m$;

for $i \leftarrow 1$ **to** m **do**

$\quad f_i = \mathscr{P}_i\left(s_t, \widehat{L}_n^{s_h^1}(f_1), \cdots, \widehat{L}_n^{s_h^{i-1}}(f_{i-1})\right)$ and compute $\widehat{L}_n^{s_h^i}(f_i)$;

[1] Oneto L, Ridella S, Anguita D. Differential privacy and generalization: Sharper bounds with applications. Pattern Recognition Letters. 2017 Apr 1;89:31-8.

Adaptive Data Analysis (II)

AS: the adaptive setting is the case when the procedures for building our models exploit both the training set and the performance of the procedure at previous step over the hold out set.

Then we need one test set at each step:

$$f_i = \mathscr{P}_i(s_t, \mathscr{P}_{i-1}, \cdots, \mathscr{P}_1)$$

$$\mathbb{P}_{\boldsymbol{S}_h^i}\left\{\exists i \in \mathcal{I}_m : \left|L(f_i) - \widehat{L}_n^{\boldsymbol{S}_h^i}(f_i)\right| \geq \sqrt{\frac{m \ln\left(\frac{2}{\delta}\right)}{2n}}\right\} \leq \delta$$

$$O\left(\sqrt{m/n}\right) \div O\left(m/n\right)$$

[1] Oneto L, Ridella S, Anguita D. Differential privacy and generalization: Sharper bounds with applications. Pattern Recognition Letters. 2017 Apr 1;89:31-8.

Thresholdout

The idea is to look at the test set error only when is far from the one on the training set, but, when we look at the error on the test set, we look at it in a “private way”.

Algorithm 1: Thresholdout for the AS

Input: s_t, s_h, T, σ, B, and $\mathscr{P}_1, \cdots, \mathscr{P}_m$
Output: $a_1, \cdots, a_m$

$\gamma \sim \mathrm{Lap}(2\sigma), \;\; \widehat{T} = T + \gamma;$
for $i \leftarrow 1$ **to** m **do**
  if $B < 1$ **then**
    $a_i = \perp;$
  else
    $f_i = \mathscr{P}_i(s_t, a_1, \cdots, a_{i-1}), \;\; \eta \sim \mathrm{Lap}(4\sigma);$
    if $|\widehat{L}_n^{s_h}(f_i) - \widehat{L}_n^{s_t}(f_i)| \geq \widehat{T} + \eta$ **then**
      $\xi \sim \mathrm{Lap}(\sigma), \;\; \gamma \sim \mathrm{Lap}(2\sigma), \;\; \widehat{T} = T + \gamma, \;\; B = B - 1;$
      $a_i = \widehat{L}_n^{s_h}(f_i) + \xi;$
    else
      $a_i = \widehat{L}_n^{s_t}(f_i);$

[1] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A., 2015c. The reusable holdout: Preserving validity in adaptive data analysis. Science 349, 636–638.
[2] Oneto L, Ridella S, Anguita D. Differential privacy and generalization: Sharper bounds with applications. Pattern Recognition Letters. 2017 Apr 1;89:31-8.

Hoeffding-type Bound

We obtain a result which is analogous to the one of the NAS setting

$$\mathbb{P}_{\boldsymbol{A}_i,\boldsymbol{F}_i}\left\{\exists i \in \mathcal{I}_m : \boldsymbol{A}_i \neq \perp \wedge |\boldsymbol{A}_i - L(\boldsymbol{F}_i)| \geq 40\sqrt{\frac{B\ln\left(\frac{12m}{\beta}\right)}{n}}\right\} \leq \beta$$

$$O\left(\sqrt{\ln(m)/n}\right)$$

[1] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A., 2015c. The reusable holdout: Preserving validity in adaptive data analysis. Science 349, 636–638.
[2] Oneto L, Ridella S, Anguita D. Differential privacy and generalization: Sharper bounds with applications. Pattern Recognition Letters. 2017 Apr 1;89:31-8.

Chernoff-type Bound

We can obtain a sharp result (at least asymptotically).

$$t = 40\sqrt{\frac{B \ln\left(\frac{12m}{\beta}\right)}{n}}$$

$$\mathbb{P}_{\boldsymbol{A}_i,\boldsymbol{F}_i}\left\{\exists i \in \mathcal{I}_m : \boldsymbol{A}_i \neq \perp \wedge |\boldsymbol{A}_i - L(\boldsymbol{F}_i)| \geq 30\sqrt{\boldsymbol{A}_i}t + 50t^2\right\} \leq \beta$$

$$O\left(\sqrt{\ln(m)/n}\right) \div O\left(\ln(m)/n\right)$$

[1] Oneto L, Ridella S, Anguita D. Differential privacy and generalization: Sharper bounds with applications. Pattern Recognition Letters. 2017 Apr 1;89:31-8.

Open Problems

- Is privacy reducing our ability to learn something from data?
- Can we improve the rate of convergence and the constants in the bounds?
- How can we exploit privacy to derive new learning algorithms?
- Can we improve the Thresholdout?
- How many times can we access the data without compromising the privacy?