PhD defense

Yiyang Yu

LPSM, Université de Paris

April 6, 2022

Advisors: Stéphane Gaïffas (LPSM, U. Paris),
Emmanuel Bacry (CEREMADE, U. Paris-Dauphine)
Rapporteurs: Alexandre Gramfort (INRIA Saclay),
Tim Van Erven (U. Amsterdam)
Examinateurs: Agathe-Chloé Azencott (Mines ParisTech),
Gilles Stoltz (U. Paris-Saclay)

Outline

Thesis title:
Deep learning in public health, and contributions to statistical learning

General background

- application of deep learning on healthcare data
- methodological research for machine learning

Contributions in this thesis

- ZiMM: a deep learning model for long term and blurry relapses with non-clinical claims data
- Contrastive learning: theoretical guarantees on the contrastive unsupervised representation learning for classification
- WildWood: a new random forest algorithm
- Online logistic regression: towards an efficient algorithm with better regret guarantee

ZiMM:
a deep learning model for long term and blurry relapses with non-clinical claims data

ZiMM: background

Deep learning and its applications

- Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, Transformer...
- applications in Computer Vision, Natural Language Processing, audio, time series, tabular data...

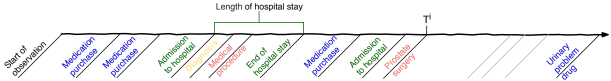
Electronic Healthcare Records (EHRs) & Claims data

- data sources: hospitals, institutions, companies... $\rightsquigarrow$ SNDS
- specificities due to the nature of healthcare data: privacy, regulations, fairness...
- other works using deep learning includes Lipton et al. (2016), Rajkomar et al. (2018), Choi et al. (2019)...

$\rightsquigarrow$ Objectives

- prediction
- representation learning
- interpretability

ZiMM: data



SNDS *(Système National des Données de Santé)*

⇝ non-clinical claims data

- 66 million French residents in total
- sequences of time-stamped events: drugs prescriptions, medical procedures, diagnostics, hospital stays
- data extraction: SCALPEL-3 (Bacry et al., 2020)

Begnin prostate hyperplasia (BPH)

- drugs regularly needed for urination problem
- post-surgical complications: persisting urination problem

⇝ TURP surgery + at least 18 months records after surgery

- 138976 patients

ZiMM: zero-inflated multinomial model

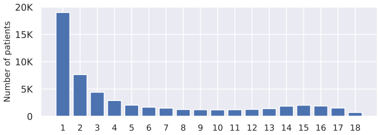


Figure: Histogram of n_i for $n_i > 0$

- Label $n_i = \sum_{b=1}^{18} y_{i,b}$, features x_i, for patient i
- We suppose

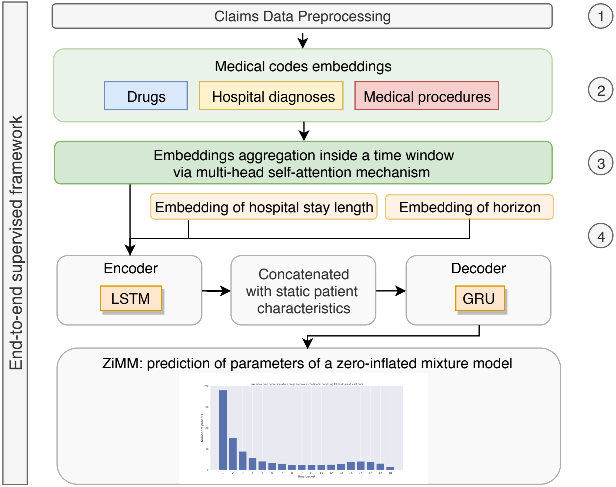
$$n_i \sim \text{Categorical}(\pi_0(x_i), \pi_1(x_i), \dots, \pi_B(x_i))$$

- and also

$$y_i | (x_i, n_i = b) \sim \begin{cases} \delta_{[0,\dots,0]} & \text{if } b = 0, \\ \text{Multinomial}(b, p_{b,1}(x_i), \dots, p_{b,B}(x_i)) & \text{otherwise} \end{cases}$$

$\rightsquigarrow$ $\pi_0(x_i), \pi_1(x_i), \dots, \pi_B(x_i)$, $p_{b,1}(x_i), \dots, p_{b,B}(x_i)$ are learnable parameters specific to patient i.

ZiMM: Encoder - Decoder architecture



ZiMM: experiments

- ZiMM Encoder-Decoder implemented in Python, Tensorflow2
- predictive performance compared to some baseline models

Model	mean-AP	AUC-ROC	AUC-PR
LRl2-SF	0.19	0.64	0.50
GBDT-SF	0.24	0.67	0.56
MLP-SF	0.18	0.64	0.49
LRl2-DF	0.21	0.65	0.53
GBDT-DF	0.25	0.68	0.57
MLP-DF	0.19	0.65	0.50
Word2vec-ISS	0.20	0.65	0.53
LSTM-ISS	0.21	0.67	0.54
Patient2Vec	-	0.68	0.55
ZiMM ED	**0.306**	**0.701**	**0.619**

- extensive experiments for hyperparameters tuning
- visualization for learned medical codes representations

ZiMM: wrap-up

- proposed a probabilistic model for blurry relapses
- exploited healthcare claims database with deep learning architectures
- possible future works:
  - including more data from hospital: vital signs, test results...
  - multi-task predictions, pre-training techniques for representation learning

Publication

ZiMM: a deep learning model for long term and blurry relapses with non-clinical claims data, Journal of Biomedical Informatics, 2020, with A. Kabeshova, B. Lukacs, S. Gaïffas, E. Bacry.

WildWood: a new random forest algorithm

WildWood: background

Setting: batch supervised learning

Random forest (Breiman, 2001)

- bootstrap aggregation: in-the-bag, out-of-bag samples
- decision tree predictor (in-the-bag + feature subsampling)

$$\widehat{f}_{(\mathcal{T},\Sigma)} : \mathcal{X} \to \widehat{\mathcal{Y}}, x \mapsto \widehat{y}_{C(x)}$$

  where $C(x)$ is the unique leaf containing x
- then RF predictor

$$\widehat{g}(\,\cdot\,;\Pi) = \frac{1}{M}\sum_{m=1}^{M} \widehat{f}(\,\cdot\,;\Pi_m)$$

Exponential weight aggregation (Cesa-Bianchi and Lugosi, 2006)

$$\widehat{f}(\cdot) := \int_\Theta \widehat{f}_\theta(\cdot) v(\mathrm{d}\theta) \quad \text{where} \quad \frac{\mathrm{d}v}{\mathrm{d}\pi}(\theta) := \frac{e^{-\eta L_\theta}}{\int_\Theta e^{-\eta L_{\theta'}} \pi(\mathrm{d}\theta')}$$

with L_θ the loss of $\widehat{f}_\theta$ on all the training data

WildWood: improved tree prediction

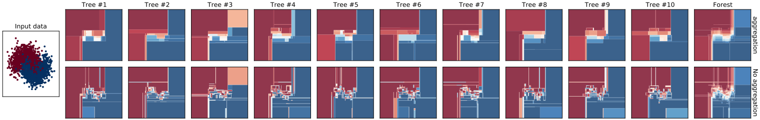
Exact aggregation

WildWood tree prediction function

$$\widehat{f}(\cdot) = \frac{\sum_{T \subset \mathcal{T}} \pi(T) e^{-\eta L_T} \widehat{y}_T(\cdot)}{\sum_{T \subset \mathcal{T}} \pi(T) e^{-\eta L_T}} \quad (1)$$

with $\eta > 0$ temperature parameter, $\pi(T) = 2^{-\|T\|}$, $\|T\|$ denotes the number of nodes in T minus its number of leaves that are also leaves of $\mathcal{T}$, $L_T := \sum_{i \in I_{\text{oob}}} \ell(\widehat{y}_T(x_i), y_i)$

- aggregation of the predictions $\widehat{y}_T(\cdot)$ of all subtrees T, weighted by $\exp(-\eta L_T)$ together with prior $\pi(T) = 2^{-\|T\|}$
- "a non-greedy way to prune trees"



Wildwood: theoretical guarantees

Theorem (Oracle inequality)

Assume that ℓ is η-exp-concave. Then, $\widehat{f}$ given by (1) *satisfies the oracle inequality*

$$\frac{1}{n_{\text{oob}}}\sum_{i \in I_{\text{oob}}} \ell(\widehat{f}(x_i), y_i) \leqslant \inf_{T \subset \mathcal{T}} \left\{ \frac{1}{n_{\text{oob}}} \sum_{i \in I_{\text{oob}}} \ell(\widehat{y}_T(x_i), y_i) + \frac{C \| T \|}{\eta(n_{\text{oob}} + 1)} \right\},$$

where the infimum is taken over all subtrees $T \subset \mathcal{T}$ rooted at `root`*, and* $C = \log 2$.

- Predictor (1) is almost as good as the best of any pruning of the same tree
- Proof inspired from Dalalyan and Tsybakov (2007)
- classification tasks + log-loss $\rightsquigarrow \eta = 1$
- regression tasks + least-square loss $\rightsquigarrow \eta = \frac{1}{8B^2}$

WildWood: efficient implementation

Theorem (Simplified version)

The prediction function (1) *can be computed recursively along the path of x in $\mathcal{T}$, with a computational cost in the same order with that of a standard RF.*

- thanks to context tree weighting (Catoni, 2004)

- native support for categorical features
- histogram strategy to accelerate split finding
- WildWood implemented in Python, open-sourced at `https://github.com/pyensemble/wildwood`

WildWood: experiments (1)

	XGB	LGBM	CB	HGB	RF10	RF100	WW10	WW100
adult	0.930	**0.931**	0.927	0.930	0.916	0.919	0.918	0.919
bank	0.933	**0.935**	0.925	0.930	0.917	0.929	0.924	0.931
breastcancer	0.991	0.993	0.987	**0.994**	0.974	0.978	0.992	0.992
car	0.999	**1.000**	**1.000**	**1.000**	0.996	0.996	0.997	0.998
covtype	**0.999**	**0.999**	0.998	**0.999**	0.996	0.998	0.996	0.998
default-cb	0.780	**0.783**	0.780	0.779	0.748	0.774	0.773	0.778
higgs	0.853	**0.857**	0.847	0.853	0.812	0.834	0.818	0.835
internet	0.934	0.910	**0.938**	0.911	0.841	0.911	0.923	0.928
kddcup	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.998	1.000	1.000
kick	**0.777**	0.770	**0.777**	0.771	0.736	0.752	0.756	0.763
letter	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.999	0.996	0.999
satimage	**0.991**	**0.991**	**0.991**	0.987	0.980	0.989	0.983	0.991
sensorless	**1.000**	**1.000**	**1.000**	**1.000**	1.000	1.000	1.000	1.000
spambase	**0.990**	**0.990**	0.987	0.986	0.980	0.986	0.983	0.987

Table: Test AUC after hyperoptimization

WildWood: experiments (2)

	Training time (seconds)						Test AUC					
	XGB	LGBM	CB	HGB	RF	WW	XGB	LGBM	CB	HGB	RF	WW
covtype	10	**3**	120	14	21	<u>3</u>	0.986	0.978	**0.989**	0.960	<u>0.998</u>	0.979
higgs	36	**30**	653	85	1389	<u>179</u>	0.823	0.812	**0.840**	0.812	<u>0.838</u>	0.813
internet	9	**4**	188	8	0.4	<u>0.3</u>	**0.918**	0.828	0.910	0.500	0.862	<u>0.889</u>
kddcup	175	41	2193	**31**	208	<u>12</u>	**1.000**	0.638	0.988	0.740	0.998	<u>1.000</u>
kick	7	**0.4**	50	0.7	31	<u>5</u>	0.768	0.757	**0.781**	0.773	0.747	<u>0.751</u>

Table: Training times default hyperparameters

WildWood: wrap-up

- improved tree prediction with exponential weight aggregation over all subtrees, at small computational cost
- efficient implementation in Python, with native support for categorical features and histogram strategy
- future directions include:
  - implementation improvement (feature importance, distributed training...)
  - subtree aggregation for other algorithms

Pre-print

WildWood: a new random forest algorithm,
with I.Merad, S. Gaïffas.

Online logistic regression: towards an efficient algorithm with better regret guarantee

Online logistic regression: background (1)

Setting: online supervised learning

Online logistic regression

At each time step $t = 1, 2, \dots,$

- environment reveals $x_t \in \mathcal{X} \subset \mathbb{R}^d$
- agent chooses $\widehat{y}_t$, based on $(x_1, y_1), \dots, (x_{t-1}, y_{t-1})$ and x_t
- environment reveals the actual $y_t \in \{-1, +1\}$
- agent suffers $\ell(\widehat{y}_t, y_t) := -\log \sigma(y_t \widehat{y}_t)$, with $\sigma(u) := \frac{1}{1+\exp(-u)}$ sigmoid

Some remarks

- log-loss with $p(y \mid x, \theta) := \sigma(y \cdot \theta^\top x)$
- Performance of successive predictions $(\widehat{y}_t)_{t=1,\dots,n}$ measured by regret

$$\mathrm{Regret}_n := \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{\theta \in \Theta} \sum_{t=1}^{n} \ell(\theta^\top x_t, y_t)$$

- Usual assumptions: with some $R, B > 0$,
  - $\|x_t\| \leqslant R$
  - comparison class $\mathcal{F} = \{x \mapsto \langle \theta, x \rangle, \|\theta\| \leqslant B\}$, $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\| \leqslant B\}$

Online logistic regression: background (2)

Algorithms for Online logistic regression

- Proper vs. improper
  - Hazan, Koren, and Levy (2014): any $O(B \log n)$ regret is impossible for proper algorithms
- non-Bayesian vs. Bayesian

Algorithm	Type	Regret upper-bound	Computational cost
OGD (Zinkevich, 2003)	Proper	$BR\sqrt{n}$	nd
ONS (Hazan, A. Agarwal, and Kale, 2007)	Proper	$de^{BR} \log(n)$	nd^2
Foster et al. (2018)	Improper, Bayesian	$d \log(BRn)$	$B^6 n^{12}(Bn + d)^{12}$
AIOLI (Jézéquel, Gaillard, and Rudi, 2020)	Improper, non-Bayesian	$dBR \log(BRn)$	$nd^2 + nBR \log(n)$
GAF* (Jézéquel, Gaillard, and Rudi, 2021)	Improper, Bayesian	$dBR \log(n) + dB^2$	$nd^2 + n^4$
FOLKLORE* (N. Agarwal, Kale, and Zimmert, 2021)	Improper, non-Bayesian	$dBR \log(n)$	$nd^2 + nBR \log(n)$
?	Improper, non-Bayesian	?$(d + B^2R^2) \log(n)$?	efficient

Sample Minmax Predictor (SMP) for batch setting

- SMP (Mourtada and Gaiffas, 2019) for logistic regression: excess loss in $O((d + B^2R^2)/n)$
- equivalent version under online setting?

OSMP: One-Step Minmax Predictor

- At time t, consider λ-ridge-penalized regret, with some $\lambda > 0$

$$\ell(\widehat{y}_t, y_t) + \underbrace{\sum_{s=1}^{t-1} \ell(\widehat{y}_s, y_s)}_{\widehat{L}_{t-1}} - \inf_{\theta \in \Theta} \Big(\ell(\theta^\top x_t, y_t) + \underbrace{\sum_{s=1}^{t-1} \ell(\theta^\top x_s, y_s) + \lambda \|\theta\|^2}_{L_{\lambda,t-1}(\theta)} \Big)$$

$\rightsquigarrow$ choosing $\widehat{y}_t$ in a minmax fashion, with $(x_s, y_s)_{s=1,\dots,t-1}$, x_t known

- OSMP uses

$$\widehat{y}_t = \operatorname*{argmin}_{\hat{y} \in \mathbb{R}} \sup_{y_t \in \{-1,1\}} \sup_{\theta \in \mathbb{R}^d} \Big\{ \ell(\widehat{y}, y_t) - \Big(\ell(\theta^\top x_t, y_t) + L_{\lambda,t-1}(\theta) \Big) \Big\}$$

- Using log-loss properties, OSMP is equivalent to

$$\widehat{y}_t = -L_{\lambda,t}^{+1\star} + L_{\lambda,t}^{-1\star}$$

with $L_{\lambda,t}^{y\star} := \inf_{\theta \in \mathbb{R}^d} \big\{ \ell(\theta^\top x_t, y) + L_{\lambda,t-1}(\theta) \big\}$ for $y \in \{-1, +1\}$

OSMP: algorithm

Algorithm 1 One-Step Minmax Predictor (OSMP) description.

Inputs: Parameters $\lambda > 0$, $n, d \geqslant 1$ constants $B, R > 0$
Initialize: $L_{\lambda,0}(\theta) = \lambda \left\|\theta\right\|^2$
for $t = 1, \ldots, n$ **do**
    Receive x_t
    Compute $L^{y\star}_{\lambda,t} \leftarrow \inf_{\theta \in \mathbb{R}^d} \left\{ \ell(\theta^\top x_t, y) + L_{\lambda,t-1}(\theta) \right\}$ for $y = -1$ and $y = +1$
    Predict $\widehat{y}_t \leftarrow -L_t^{+1\star} + L_t^{-1\star}$
    Receive y_t
    Update function $L_{\lambda,t}(\theta) = L_{\lambda,t-1}(\theta) + \ell(\theta^\top x, y_t)$
end for

OSMP: regret analysis

- Consider $\mathrm{Regret}_n(\theta)$ relative to a comparison parameter θ, then

$$\mathrm{Regret}_n(\theta) \leqslant \sum_{t=1}^{n} \underbrace{\left(\ell(\widehat{y}_t, y_t) - \inf_{\theta' \in \mathbb{R}^d} L_{\lambda,t}(\theta') + \inf_{\theta' \in \mathbb{R}^d} L_{\lambda,t-1}(\theta') \right)}_{\widehat{r}_t} + \lambda \|\theta\|^2$$

- Assume $\|x_t\| \leqslant R$. Running OSMP with $\lambda \geqslant R^2$, then

$$\widehat{r}_t \leqslant e \cdot \sigma'(\langle \theta_t, x_t \rangle) \cdot \|x_t\|^2_{(\nabla^2 L_{\lambda,t}(\theta_t))^{-1}}$$

  with $\theta_t := \mathrm{argmin}_{\theta \in \mathbb{R}^d} L_{\lambda,t}(\theta)$
  - generalized self-concordance of $L_{\lambda,t}$
- But

$$\sum_{t=1}^{n} \sigma'(\langle \theta_t, x_t \rangle) \cdot \|x_t\|^2_{(\nabla^2 L_{\lambda,t}(\theta_t))^{-1}} \leqslant ?$$

AOSMP: Approximated One-Step Minmax Predictor

- AOSMP uses

$$\widehat{y}_t = \operatorname*{argmin}_{\widehat{y}\in\mathbb{R}} \sup_{y_t\in\{-1,+1\}} \sup_{\theta\in\mathbb{R}^d} \left\{ \ell(\widehat{y}, y_t) - \left(\ell(\theta^\top x_t, y_t) + \tilde{L}_{\lambda,t-1}(\theta) \right) \right\}$$

with $\tilde{L}_{\lambda,t-1}(\theta) = \sum_{s=1}^{t-1} \tilde{\ell}(\theta^\top x_s, y_s) + \lambda \|\theta\|^2$, and

$$\tilde{\ell}_t(\theta) := \ell_t(\tilde{\theta}_t) + g_t^\top(\theta - \tilde{\theta}_t) + \frac{1}{2}\eta_t \left(x_t^\top(\theta - \tilde{\theta}_t) \right)^2, \qquad (2)$$

with $g_t := \nabla \ell_t(\tilde{\theta}_t)$, $\eta_t := \sigma'(\langle \tilde{\theta}_t, x_t \rangle)/(1 + BR)$ and we choose

$$\tilde{\theta}_t = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \tilde{L}^{y_t}_{\lambda,t}(\theta) \quad \text{with} \quad \tilde{L}^{y}_{\lambda,t}(\theta) := \ell(\theta^\top x_t, y) + \tilde{L}_{\lambda,t-1}(\theta)$$

- Inequality from AIOLI (Jézéquel, Gaillard, and Rudi, 2020): for any $\|\theta\| \leqslant B$,

$$\tilde{\ell}_t(\theta) \leqslant \ell_t(\theta)$$

- Using log-loss properties,

$$\widehat{y}_t = -\tilde{L}^{+1\star}_{\lambda,t} + \tilde{L}^{-1\star}_{\lambda,t}$$

with $\tilde{L}^{y\star}_{\lambda,t} := \inf_{\theta\in\mathbb{R}^d} \left\{ \ell(\theta^\top x_t, y) + \tilde{L}_{\lambda,t-1}(\theta) \right\}$ for $y \in \{-1, +1\}$

AOSMP: algorithm

Algorithm 2 Approximated One-Step Minmax Predictor (AOSMP) description.

Inputs: Parameters $\lambda > 0$, $n, d \geqslant 1$, constants $B, R > 0$
Initialize: Function $\tilde{L}_{\lambda,0}(\theta) = \lambda \|\theta\|^2$
for $t = 1, \ldots, n$ **do**
  Receive x_t
  Compute $\tilde{L}^{y\star}_{\lambda,t} \leftarrow \inf_{\theta \in \mathbb{R}^d} \left\{ \ell(\theta^\top x_t, y) + \tilde{L}_{\lambda,t-1}(\theta) \right\}$ for $y = -1$ and $y = +1$
  Predict $\widehat{y}_t \leftarrow -\tilde{L}^{+1\star}_{\lambda,t} + \tilde{L}^{-1\star}_{\lambda,t}$
  Receive y_t
  Compute $\tilde{\theta}_t \leftarrow \mathrm{argmin}_{\theta \in \mathbb{R}^d} \left\{ \ell(\theta^\top x_t, y_t) + \tilde{L}_{\lambda,t-1}(\theta) \right\}$, and function $\tilde{\ell}_t(\theta)$ as specified in Equation (2)
  Update function $\tilde{L}_{\lambda,t}(\theta) = \tilde{L}_{\lambda,t-1}(\theta) + \tilde{\ell}_t(\theta)$
end for

AOSMP: regret analysis

Theorem

Let $(x_1, y_1), \ldots, (x_n, y_n) \in [-R, R]^d \times \{-1, +1\}$ be an arbitrary sequence of observations. We run AOSMP (Algorithm 2) with $\lambda \geqslant R^2$. Then its regret against any $\|\theta\| \leqslant B$ satisfies

$$\mathrm{Regret}_n(\theta) \leqslant e \cdot (1 + BR) d \log \left(1 + \frac{nR^2}{8d(1+BR)\lambda}\right) + \lambda \|\theta\|^2 .$$

In particular, choosing $\lambda = R^2$ yields

$$\mathrm{Regret}_n \leqslant e \cdot (1 + BR) d \log \left(1 + \frac{n}{8d(1+BR)}\right) + B^2 R^2.$$

- pseudo instant regret $\widehat{r}_t := \ell(\widehat{y}_t, y_t) - \tilde{L}^\star_{\lambda,t} + \tilde{L}^\star_{\lambda,t-1}$
- again generalized self-concordance of $\tilde{L}_t$,

$$\widehat{r}_t \leqslant e \cdot \sigma'(\langle \tilde{\theta}_t, x_t \rangle) \cdot \left\langle (\nabla^2 \tilde{L}^{y_t}_{\lambda,t}(\tilde{\theta}_t))^{-1} x_t, x_t \right\rangle$$

with $\tilde{\theta}_t = \mathrm{argmin}_{\theta \in \mathbb{R}^d} \tilde{L}^{y_t}_{\lambda,t}(\theta)$, and

$$\nabla^2 \tilde{L}^{y_t}_{\lambda,t}(\tilde{\theta}_t) = \sigma'(\langle \tilde{\theta}_t, x_t \rangle) x_t x_t^\top + \sum_{s=1}^{t-1} \frac{1}{1+BR} \sigma'(\langle \tilde{\theta}_s, x_s \rangle) x_s x_s^\top + 2\lambda I_d$$

OSMP, AOSMP: wrap-up

- One-Step Minmax Predictor for online logistic regression
  - OSMP
  - AOSMP: using approximated loss
- regret analysis

Algorithm	Type	Regret upper-bound	Computational cost
OGD	Proper	$BR\sqrt{n}$	nd
ONS	Proper	$de^{BR}\log(n)$	nd^2
Foster	Improper, Bayesian	$d\log(BRn)$	$B^6 n^{12}(Bn+d)^{12}$
GAF*	Improper, Bayesian	$dBR\log(n) + dB^2$	$nd^2 + n^4$
AIOLI	Improper, non-Bayesian	$dBR\log(BRn)$	$nd^2 + nBR\log(n)$
FOLKLORE*	Improper, non-Bayesian	$dBR\log(n)$	$nd^2 + nBR\log(n)$
?	Improper, non-Bayesian	$?(d + B^2R^2)\log(n)?$	efficient
AOSMP	Improper, non-Bayesian	$dBR\log(n) + B^2R^2$	$nd^2 + nC_{GD}$

Collaboration with I. Merad, J. Mourtada, S. Gaïffas.

Conclusion

Concluding remarks

In this presentaion

- a deep learning model for post-surgical complications prediciton with non-clinical claims data (ZiMM)
- A new random forest algorithm (WildWood)
- Studies on some new algorithms for online logistic regression (OSMP, AOSMP)

Thank you for your attention!

Bibliography I

Agarwal, Naman, Satyen Kale, and Julian Zimmert (2021). "Efficient Methods for Online Multiclass Logistic Regression". In: *arXiv preprint arXiv:2110.03020*.

Bacry, Emmanuel et al. (2020). "SCALPEL3: a scalable open-source library for healthcare claims databases". In: *International Journal of Medical Informatics*, p. 104203.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Catoni, Olivier (2004). *Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001*. Vol. 1851. Lecture Notes in Mathematics. Springer-Verlag Berlin Heidelberg. ISBN: 9783540445074. URL: https://books.google.fr/books?id=3Ih8CwAAQBAJ.

Cesa-Bianchi, Nicolo and Gábor Lugosi (2006). *Prediction, learning, and games*. Cambridge university press.

Choi, Edward et al. (June 2019). "Graph Convolutional Transformer: Learning the Graphical Structure of Electronic Health Records". In: arXiv: 1906.04716. URL: http://arxiv.org/abs/1906.04716.

Bibliography II

Dalalyan, Arnak S and Alexandre B Tsybakov (2007). "Aggregation by exponential weighting and sharp oracle inequalities". In: *International Conference on Computational Learning Theory*. Springer, pp. 97–111.

Foster, Dylan J. et al. (June 2018). "Logistic Regression: The Importance of Being Improper". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, pp. 167–208. URL: http://proceedings.mlr.press/v75/foster18a.html.

Hazan, Elad, Amit Agarwal, and Satyen Kale (2007). "Logarithmic regret algorithms for online convex optimization". In: *Machine Learning* 69.2-3, pp. 169–192.

Hazan, Elad, Tomer Koren, and Kfir Y Levy (2014). "Logistic regression: Tight bounds for stochastic and online optimization". In: *Conference on Learning Theory*. PMLR, pp. 197–209.

Jézéquel, Rémi, Pierre Gaillard, and Alessandro Rudi (2020). "Efficient improper learning for online logistic regression". In: *Conference on Learning Theory*. PMLR, pp. 2085–2108.

Bibliography III

Jézéquel, Rémi, Pierre Gaillard, and Alessandro Rudi (Oct. 2021). "Mixability made efficient: Fast online multiclass logistic regression". working paper or preprint. URL: https://hal.archives-ouvertes.fr/hal-03370530.

Lipton, Zachary C. et al. (Nov. 2016). "Learning to diagnose with LSTM recurrent neural networks". In: *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*. arXiv: 1511.03677. URL: http://arxiv.org/abs/1511.03677.

Mourtada, Jaouad and Stephane Gaiffas (2019). "An improper estimator with optimal excess risk in misspecified density estimation and logistic regression". In: *arXiv preprint arXiv:1912.10784*.

Rajkomar, Alvin et al. (Jan. 2018). "Scalable and accurate deep learning with electronic health records". In: *npj Digital Medicine* 1.1. ISSN: 2398-6352. DOI: 10.1038/s41746-018-0029-1. arXiv: 1801.07860. URL: https://arxiv.org/abs/1801.07860.

Zinkevich, Martin (2003). "Online convex programming and generalized infinitesimal gradient ascent". In: *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936.