Contributions to statistical learning: Density estimation, expert aggregation and random forests

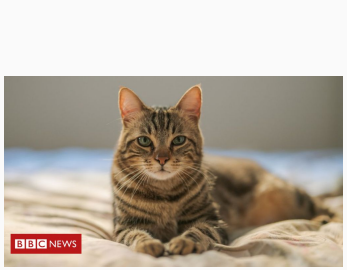
Jaouad Mourtada

Supervised by Stéphane Gaïffas and Erwan Scornet

June 8th, 2020

Supervised Statistical Learning

Predict a quantity of interest (**label**, output) based on associated variables (input), given some **examples**



`cat, dog`

Inputs **Outputs**

Applications: visual object recognition (e.g. tumor detection), speech recognition, automatic translation...

Principle: use available data to find **correlations** between inputs and outputs

Overview of the thesis

1. **Mondrian Random forests**
   - Chap. 2: statistical analysis [M., Gaïffas, Scornet 2018]
   - Chap. 3: efficient online version [M., Gaïffas, Scornet 2019]
2. **Expert aggregation**
   - Chap. 4: behavior of Hedge in stochastic regime [M., Gaïffas 2019]
   - Chap. 5: tracking growing expert classes [M., Maillard 2017]
3. **Density estimation, least squares and logistic regression**
   - Chap. 6: linear least squares and covariance matrices [M. 2019]
   - Chap. 7: density estimation and logistic regression [M., Gaïffas 2019]

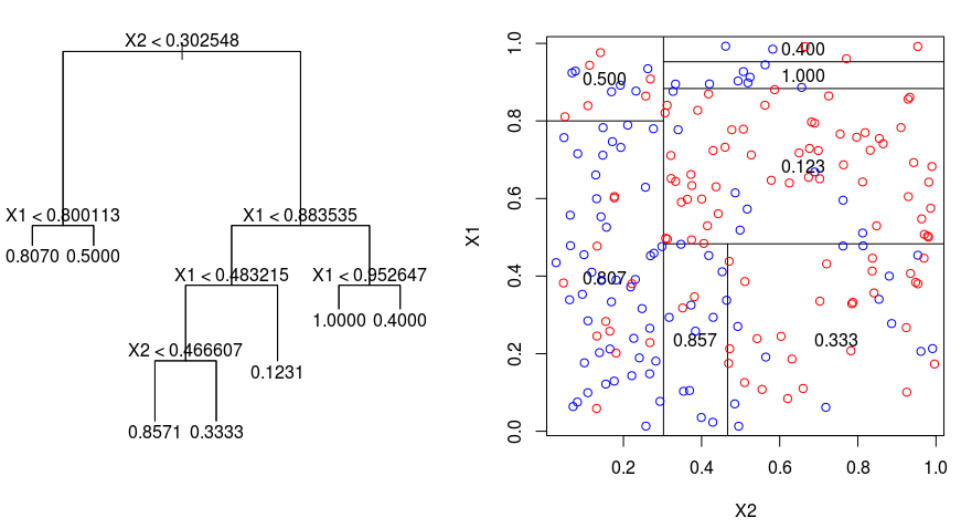
Overview of the thesis

1. **Mondrian Random forests**
   - Chap. 2: statistical analysis [M., Gaïffas, Scornet 2018]
   - Chap. 3: efficient online version [M., Gaïffas, Scornet 2019]
2. **Expert aggregation**
   - Chap. 4: behavior of Hedge in stochastic regime [M., Gaïffas 2019]
   - Chap. 5: tracking growing expert classes [M., Maillard 2017]
3. **Density estimation, least squares and logistic regression**
   - Chap. 6: linear least squares and covariance matrices [M. 2019]
   - Chap. 7: density estimation and logistic regression [M., Gaïffas 2019]

Most of this presentation will be about the last part

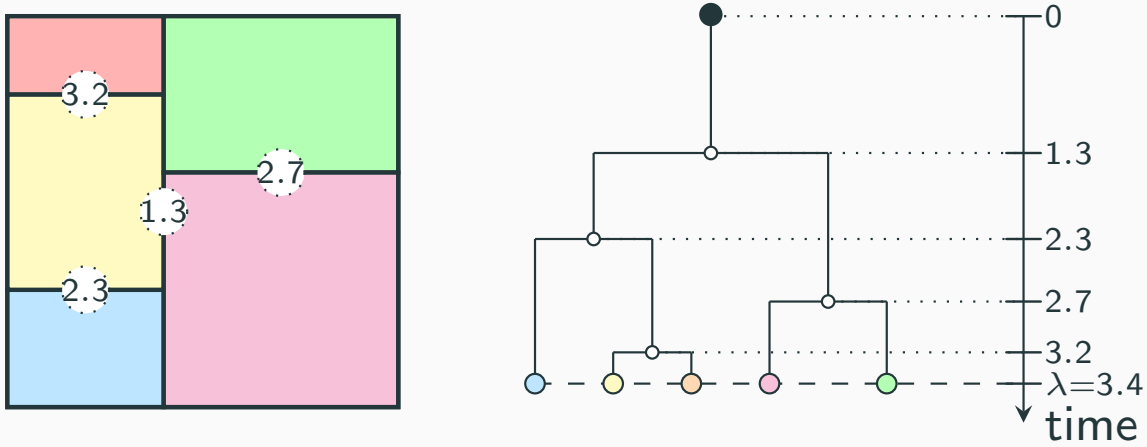
Mondrian Random forests

Random forests



- **Decision tree** partitions space recursively with axis-aligned splits, then **constant prediction** on each cell
- **Random forests** average forecasts of **randomized** decision trees
- Introduced by Breiman (2001), **often used** in classification and regression

Mondrian Random Forests [M., Gaïffas, Scornet; 2018, 2019]



A variant of Random Forests introduced by Roy and Teh (2008)

Analytically tractable: **exact distribution of cells**

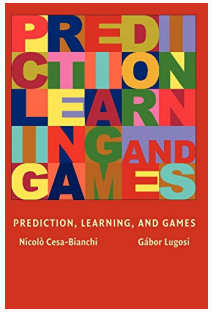
Unlike other simplified RF, achieve **minimax nonparametric rates**

Effect of averaging, extends Arlot & Genuer 2014 for 1d forests

Efficient **online** implementation

Prediction with expert advice

Prediction with expert advice[1,2]



Experts $1 \leqslant i \leqslant M$: sources of predictions at time $1 \leqslant t \leqslant T$

Sequentially combine forecasts, predict as well as best one

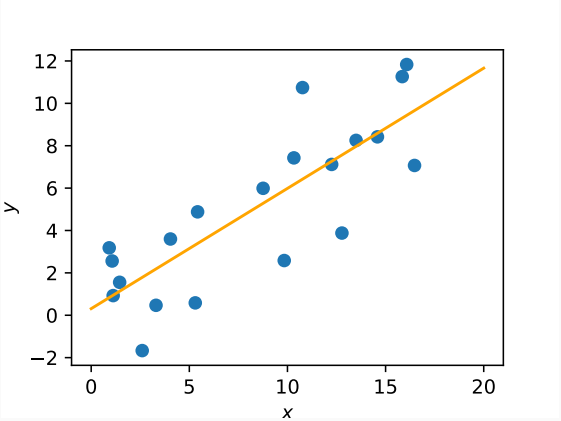
Standard **Exponential Weights** algorithm: **worst-case optimal**

Beyond worst case[2]: characterize behavior of simple variants on stochastic problems, to identify benefit of adaptive algorithms

[1] [M., Maillard, 2017]; [2] [M., Gaïffas, 2019]

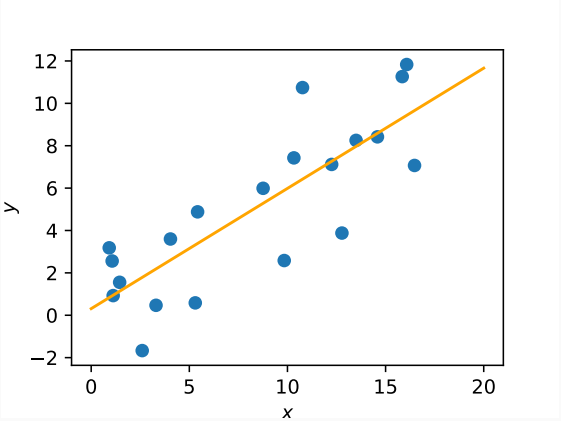
Least squares regression and covariance matrices

Linear regression



- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$; $R(\theta) = \mathbb{E}[(Y - \langle \theta, X \rangle)^2]$ risk of $\theta \in \mathbb{R}^d$

Linear regression

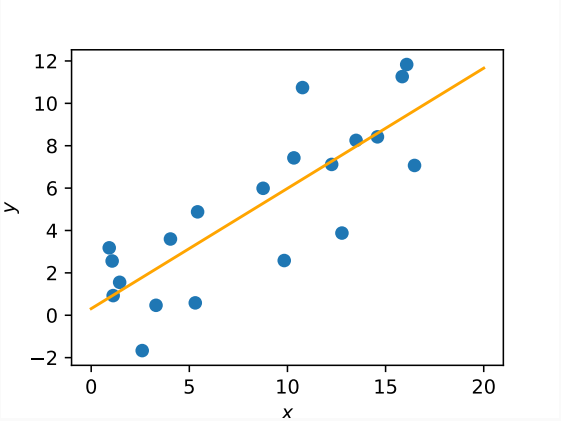


- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$; $R(\theta) = \mathbb{E}[(Y - \langle \theta, X \rangle)^2]$ risk of $\theta \in \mathbb{R}^d$
- $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. sample
- **Goal**: find $\widehat{\theta}_n$ with small **excess risk**

$$\mathcal{E}(\widehat{\theta}_n) := R(\widehat{\theta}_n) - \inf_{\theta \in \mathbb{R}^d} R(\theta) = \|\widehat{\theta}_n - \theta^*\|_\Sigma^2$$

with $\Sigma := \mathbb{E}[XX^\top]$ covariance matrix and $\theta^* := \Sigma^{-1}\mathbb{E}[YX]$

Linear regression



- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$; $R(\theta) = \mathbb{E}[(Y - \langle \theta, X \rangle)^2]$ risk of $\theta \in \mathbb{R}^d$
- $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. sample
- **Goal**: find $\widehat{\theta}_n$ with small **excess risk**

$$\mathcal{E}(\widehat{\theta}_n) := R(\widehat{\theta}_n) - \inf_{\theta \in \mathbb{R}^d} R(\theta) = \|\widehat{\theta}_n - \theta^*\|_\Sigma^2$$

  with $\Sigma := \mathbb{E}[XX^\top]$ covariance matrix and $\theta^* := \Sigma^{-1}\mathbb{E}[YX]$
- No prior knowledge on $\theta^* \in \mathbb{R}^d$
- Hardness as a function of distribution P_X of X

Minimax excess risk

$Y = \langle \theta^*, X \rangle + \varepsilon$ with $\mathbb{E}[\varepsilon X] = 0$. $P_{(X,Y)}$ depends on $P_X, P_{\varepsilon|X}, \theta^*$

Minimax excess risk

$Y = \langle \theta^*, X \rangle + \varepsilon$ with $\mathbb{E}[\varepsilon X] = 0$. $P_{(X,Y)}$ depends on $P_X, P_{\varepsilon|X}, \theta^*$

- P_X **fixed**;
- $\mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2$ and $\mathbb{E}[\varepsilon|X] = 0$ (well-specified);
- $\theta^* \in \mathbb{R}^d$ **arbitrary**

Minimax excess risk

$Y = \langle \theta^*, X \rangle + \varepsilon$ with $\mathbb{E}[\varepsilon X] = 0$. $P_{(X,Y)}$ depends on $P_X, P_{\varepsilon|X}, \theta^*$

- P_X **fixed**;
- $\mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2$ and $\mathbb{E}[\varepsilon|X] = 0$ (well-specified);
- $\theta^* \in \mathbb{R}^d$ **arbitrary**

Minimax excess risk

$$\mathcal{E}_n^*(P_X, \sigma^2) = \inf_{\widehat{\theta}_n} \sup_{P} \mathbb{E}[R(\widehat{\theta}_n) - R(\theta^*)]$$

Minimax risk in the well-specified case

P_X **degenerate** if $\mathbb{P}(X \in H) > 0$ for some hyperplane $H \subset \mathbb{R}^d$

Minimax risk in the well-specified case

P_X **degenerate** if $\mathbb{P}(X \in H) > 0$ for some hyperplane $H \subset \mathbb{R}^d$

Proposition

If P_X is **degenerate** *or $n < d$, minimax excess risk is infinite:*

$$\inf_{\widehat{\theta}_n} \sup_{P \in \mathcal{P}(P_X, \sigma^2)} \mathbb{E}[\mathcal{E}(\widehat{\theta}_n)] = +\infty$$

Minimax risk in the well-specified case

P_X **degenerate** if $\mathbb{P}(X \in H) > 0$ for some hyperplane $H \subset \mathbb{R}^d$

Proposition

If P_X is **degenerate** *or $n < d$, minimax excess risk is infinite:*

$$\inf_{\widehat{\theta}_n} \sup_{P \in \mathcal{P}(P_X, \sigma^2)} \mathbb{E}[\mathcal{E}(\widehat{\theta}_n)] = +\infty$$

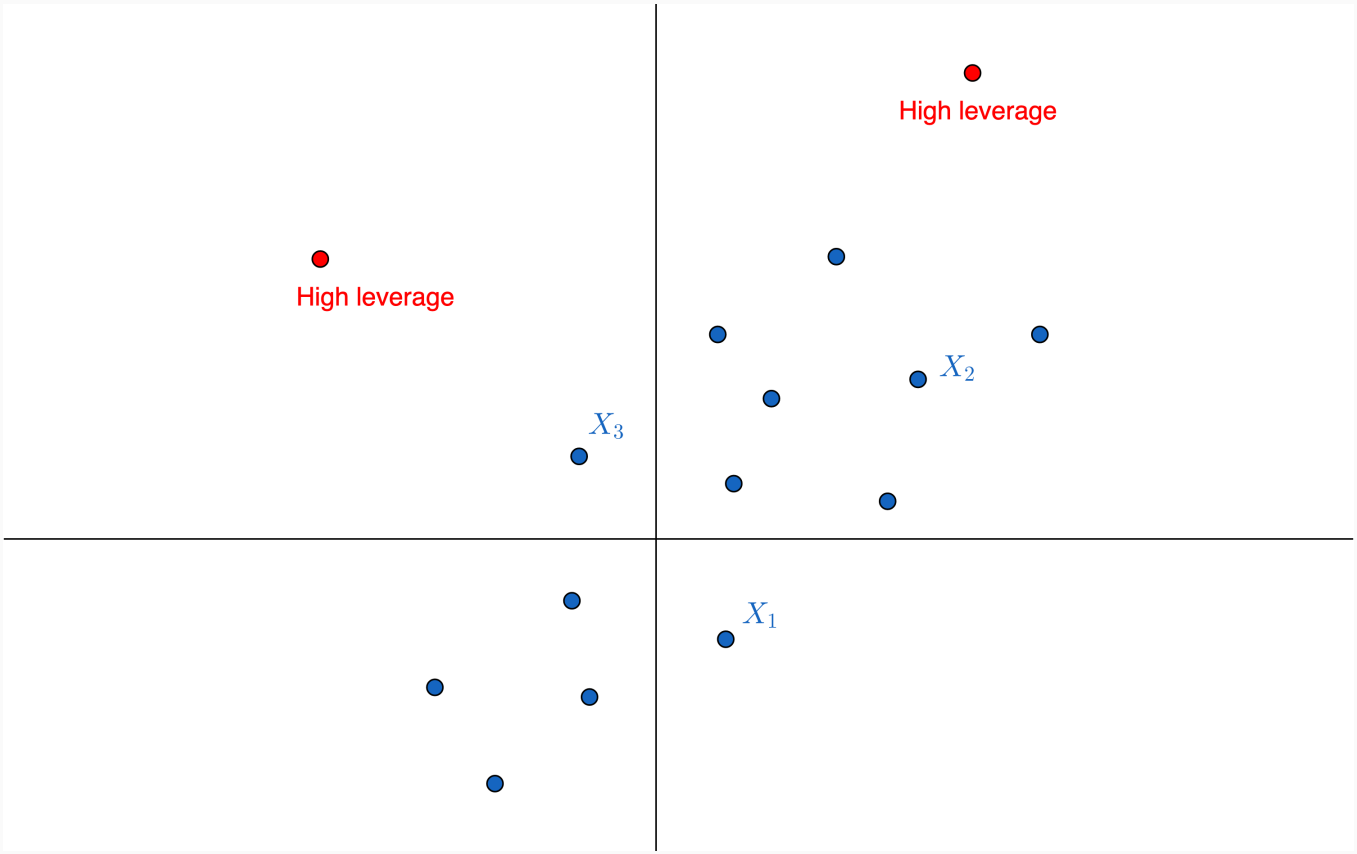
Otherwise, OLS estimator is **uniquely defined** *and* **minimax**

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}^{\mathbf{LS}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (Y_i - \langle \theta, X_i \rangle)^2$$

Link with leverage scores

Leverage score of X_{n+1} among $(X_i)_{i=1}^{n+1}$ $[\widehat{Y}_{n+1} = \langle \widehat{\theta}^{LS}_{n+1}, X_{n+1} \rangle]$

$$\widehat{\ell}_{n+1} := \frac{\partial \widehat{Y}_{n+1}}{\partial Y_{n+1}} = \left\langle \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} X_{n+1}, X_{n+1} \right\rangle$$



Link with leverage scores

Leverage score of X_{n+1} among $(X_i)_{i=1}^{n+1}$ $[\widehat{Y}_{n+1} = \langle \widehat{\theta}^{LS}_{n+1}, X_{n+1} \rangle]$

$$\widehat{\ell}_{n+1} := \frac{\partial \widehat{Y}_{n+1}}{\partial Y_{n+1}} = \left\langle \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} X_{n+1}, X_{n+1} \right\rangle$$

Theorem (M., 2019)

Minimax excess risk over $\mathcal{P}(P_X, \sigma^2)$ *is* **characterized by distribution of leverage scores** $\widehat{\ell}_{n+1}$*: for every* P_X*,*

$$\inf_{\widehat{\theta}_n} \sup_{P \in \mathcal{P}(P_X, \sigma^2)} \mathbb{E}[\mathcal{E}(\widehat{\theta}_n)] = \sigma^2 \cdot \mathbb{E}\left[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}} \right]$$

Uneven leverage scores $\Rightarrow$ $\widehat{\theta}^{\mathrm{LS}}_n$ effectively determined by less points $\Rightarrow$ **higher minimax risk**

Link with leverage scores

Leverage score of X_{n+1} among $(X_i)_{i=1}^{n+1}$ $[\widehat{Y}_{n+1} = \langle \widehat{\theta}^{LS}_{n+1}, X_{n+1} \rangle]$

$$\widehat{\ell}_{n+1} := \frac{\partial \widehat{Y}_{n+1}}{\partial Y_{n+1}} = \left\langle \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} X_{n+1}, X_{n+1} \right\rangle$$

Theorem (M., 2019)

Minimax excess risk over $\mathcal{P}(P_X, \sigma^2)$ *is* **characterized by distribution of leverage scores** $\widehat{\ell}_{n+1}$*: for every* P_X*,*

$$\inf_{\widehat{\theta}_n} \sup_{P \in \mathcal{P}(P_X, \sigma^2)} \mathbb{E}[\mathcal{E}(\widehat{\theta}_n)] = \sigma^2 \cdot \mathbb{E}\left[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}} \right] \geqslant \frac{\sigma^2 d}{n - d + 1}$$

Uneven leverage scores $\Rightarrow$ $\widehat{\theta}_n^{\mathrm{LS}}$ effectively determined by less points $\Rightarrow$ **higher minimax risk**

Distribution-dependent lower bound

- For **every** P_X, minimax risk $\geqslant \sigma^2 d/(n-d+1)$
- If $n, d \to \infty$, $d/n \to \gamma \in (0,1)$: $\boldsymbol{\sigma^2\gamma/(1-\gamma)}$ **lower bound**

Distribution-dependent lower bound

- For **every** P_X, minimax risk $\geqslant \sigma^2 d/(n-d+1)$
- If $n, d \to \infty$, $d/n \to \gamma \in (0,1)$: $\boldsymbol{\sigma^2\gamma/(1-\gamma)}$ **lower bound**
- Random-design regression **harder than fixed-design regression** (in minimax sense): there $\sigma^2 d/n \to \boldsymbol{\sigma^2 \cdot \gamma}$

Distribution-dependent lower bound

- For **every** P_X, minimax risk $\geqslant \sigma^2 d/(n-d+1)$
- If $n, d \to \infty$, $d/n \to \gamma \in (0,1)$: $\boldsymbol{\sigma^2\gamma/(1-\gamma)}$ **lower bound**
- Random-design regression **harder than fixed-design regression** (in minimax sense): there $\sigma^2 d/n \to \boldsymbol{\sigma^2 \cdot \gamma}$
- If $X \sim \mathcal{N}(0, \Sigma)$ Gaussian, $\sigma^2 d/(n-d-1) \to \sigma^2\gamma/(1-\gamma)$
- Almost easiest design, constant leverage in high dimension
- Complements "universality" results for **independent** covariates

Dependence on signal strength

$Y|X \sim \mathcal{N}(\langle \theta^*, X \rangle, \sigma^2)$. **Prior** $\theta^* \sim \mathcal{N}(0, (\sigma^2\eta^2/d)\Sigma^{-1})$.
$\eta^2 = \mathbb{E}[\langle \theta^*, X \rangle^2]/\sigma^2$ **signal-to-noise ratio** (SNR)

Theorem ("Marchenko-Pastur" lower bound; M., 2019)

For any distribution P_X with $\mathbb{E}[XX^\top] = \Sigma$, Bayes risk larger than

$$\sigma^2 \cdot \frac{-\left(n+1-d+d/\eta^2\right) + \sqrt{(n+1-d+d/\eta^2)^2 + 4d^2/\eta^2}}{2d/\eta^2}.$$

- Higher than fixed-design risk $\sigma^2 \frac{\eta^2 d/n}{\eta^2 + d/n}$
- **Matching limit** for Gaussian design as $d/n \to \gamma$ (Dicker 2016):
  $\sigma^2 \cdot \frac{-(1-\gamma+\gamma/\eta^2) + \sqrt{(1-\gamma+\gamma/\eta^2)^2 + 4\gamma^2/\eta^2}}{2\gamma/\eta^2}$

Minimax upper bound: small-ball property

Requires to control **negative moments** of $\widehat{\Sigma}_n = n^{-1} \sum_{i=1}^n X_i X_i^\top$

Key assumption: **"small ball" property**[2]: $\exists C \geqslant 1, \alpha \in (0,1)$ s.t.

$$\forall t > 0, H \text{ hyperplane}, \quad \mathbb{P}(\text{dist}(\Sigma^{-1/2} X, H) \leqslant t) \leqslant (Ct)^\alpha \qquad (1)$$

[2] Koltchinskii & Mendelson 2015, Mendelson 2014, Lecué & Mendelson 2016 (for a single $t \in (0, C^{-1})$).

Minimax upper bound: small-ball property

Requires to control **negative moments** of $\widehat{\Sigma}_n = n^{-1} \sum_{i=1}^n X_i X_i^\top$

Key assumption: **"small ball" property**: $\exists C \geqslant 1, \alpha \in (0,1)$ s.t.

$$\forall t > 0, H \text{ hyperplane}, \quad \mathbb{P}(\operatorname{dist}(\Sigma^{-1/2} X, H) \leqslant t) \leqslant (Ct)^\alpha \qquad (1)$$

Theorem (M., 2019)

Under assumption (1), there exist C', c' such that for $n \gtrsim d$

$$\forall t \in (0,1), \quad \mathbb{P}(\lambda_{\min}(\Sigma^{-1/2} \widehat{\Sigma}_n \Sigma^{-1/2}) \leqslant t) \leqslant (C' t)^{c' n}$$

This is **unimprovable** *for $t \in (0, c), d \geqslant 2$; and (1) is necessary.*

Relies on **PAC-Bayesian** technique; here $t \in (0, c)$, complements results of (Oliveira 2016, Koltchinskii & Mendelson 2015) for $t \in (c, 1)$

Minimax upper bound

Small-ball: $\operatorname{dist}(\Sigma^{-1/2}X, H) \leqslant (Ct)^{\alpha}$ for all $t > 0$

Kurtosis: $\mathbb{E}\|\Sigma^{-1/2}X\|^4 \leqslant \kappa d^2$

Theorem (Risk of OLS: well-specified case)

Under those assumptions, if $P \in \mathcal{P}(P_X, \sigma^2)$, and $n \geqslant 6d/\alpha$,

$$\frac{\sigma^2 d}{n} \leqslant \mathbb{E}[\mathcal{E}(\widehat{\theta}_n^{\mathrm{LS}})] \leqslant \frac{\sigma^2 d}{n}\left(1 + C'\kappa\frac{d}{n}\right)$$

where $C' = 28C^4 e^{1+9/\alpha}$.

Bound in **expectation**; previous results for OLS in probability only

Also bound in **misspecified case** under 4th moment assumption

Summary

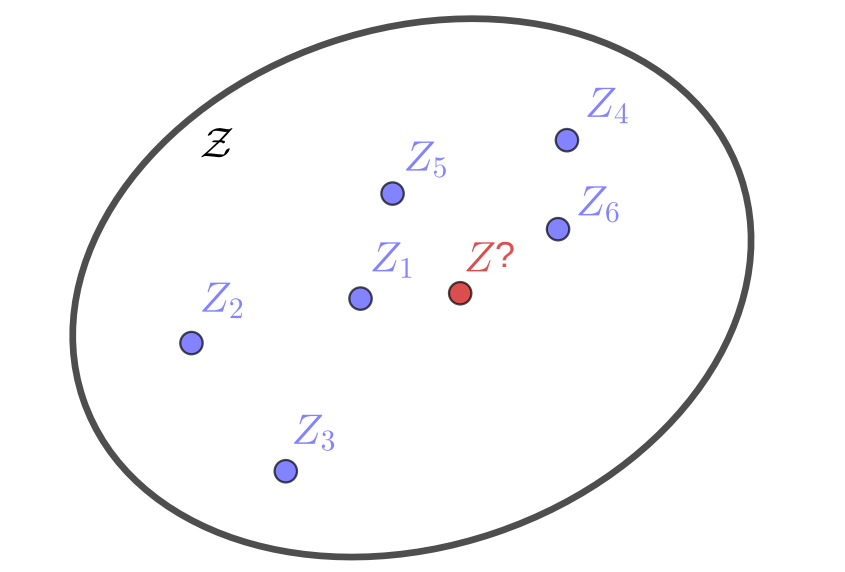
- Random-design linear regression
- Minimax risk characterized by **leverage scores**
- **Lower bound** achieved by **Gaussian design**
- **Upper bounds** based on study of empirical covariance $\widehat{\Sigma}_n$

Misspecified density estimation and logistic regression

Predictive density estimation

Predictive density estimation: setting

- Sample $Z_1^n = (Z_1, \ldots, Z_n) \sim P^n$ on $\mathcal{Z}^n$, P **unknown**
- **Predict** new sample $Z \sim P$ (**probabilistic** prediction)
- f density, point z: **log-loss** $\ell(f, z) = -\log f(z)$. Risk $R(f) = \mathbb{E}[\ell(f, Z)]$



On logarithmic loss: $\ell(f, z) = -\log f(z)$

- Standard loss function, connected to lossless compression

On logarithmic loss: $\ell(f, z) = -\log f(z)$

- Standard loss function, connected to lossless compression
- Minimizing R amounts to **maximizing joint probability** given to a large test sample $(Z'_1, \ldots, Z'_m) \sim P^m$:

$$\prod_{i=1}^{m} f(Z'_i) = \exp\Big(-\sum_{i=1}^{m} \ell(f, Z'_i) \Big) = \exp\big[-m(R(f) + o(1)) \big]$$

On logarithmic loss: $\ell(f, z) = -\log f(z)$

- Standard loss function, connected to lossless compression
- Minimizing R amounts to **maximizing joint probability** given to a large test sample $(Z'_1, \ldots, Z'_m) \sim P^m$:

$$\prod_{i=1}^{m} f(Z'_i) = \exp\Big(-\sum_{i=1}^{m} \ell(f, Z'_i)\Big) = \exp\big[-m(R(f) + o(1))\big]$$

- Risk minimized by **true density** $p = dP/d\mu$, and

$$R(f) - R(p) = \mathbb{E}_{Z\sim P}\Big[\log\Big(\frac{p(Z)}{f(Z)}\Big)\Big] = \mathrm{KL}(p, f) \geqslant 0$$

is the **Kullback-Leibler divergence** (relative entropy)

Well-specified case: asymptotic optimality of the MLE

$\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ regular parametric model of **dimension** d.
Assume $p \in \mathcal{F}$ (**well-specified** model).

The **Maximum Likelihood Estimator (MLE)** (or ERM) $\widehat{f}_n$:

$$\widehat{f}_n := \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i) = \operatorname*{argmax}_{f \in \mathcal{F}} \prod_{i=1}^{n} f(Z_i)$$

satisfies, as $n \to \infty$, (e.g., van der Vaart 1998)

$$\mathbb{E}[R(\widehat{f}_n)] - \inf_{f \in \mathcal{F}} R(f) = \frac{d}{2n} + o\left(\frac{1}{n}\right) .$$

$d/(2n)$ rate is **asymptotically optimal**: MLE is **efficient**.

Misspecified case (agnostic statistical learning)

Assumption $p \in \mathcal{F}$ is **restrictive** and generally not satisfied

General **misspecified case** where $p \notin \mathcal{F}$: model $\mathcal{F}$ is **wrong but useful**. **Excess risk**

$$\mathcal{E}(\widehat{f}_n) := R(\widehat{f}_n) - \inf_{f \in \mathcal{F}} R(f)$$

Misspecified case (agnostic statistical learning)

Assumption $p \in \mathcal{F}$ is **restrictive** and generally not satisfied

General **misspecified case** where $p \notin \mathcal{F}$: model $\mathcal{F}$ is **wrong but useful**. **Excess risk**

$$\mathcal{E}(\widehat{f}_n) := R(\widehat{f}_n) - \inf_{f \in \mathcal{F}} R(f)$$

MLE $\widehat{f}_n$ can degrade under model misspecification:

$$\mathbb{E}[R(\widehat{f}_n)] - \inf_{f \in \mathcal{F}} R(f) = \frac{d_{\text{eff}}}{2n} + o\left(\frac{1}{n}\right)$$

($d_{\text{eff}} = \text{tr}[H^{-1/2} G H^{-1/2}]$, $G = \mathbb{E}[\nabla \ell(\theta^*, Z) \nabla \ell(\theta^*, Z)^\top]$, $H = \nabla^2 R(\theta^*)$)

d_{eff} **depends on** P, and we may have $d_{\text{eff}} \gg d$.

Cumulative risk and online-to-batch conversion

Well-understood **sequential** problem (Merhav 1998, Cesa-Bianchi & Lugosi 2006): there exist $\widehat{g}_0, \dots, \widehat{g}_{n-1}$ s.t., for every $Z_1, \dots, Z_n$,

$$\sum_{t=1}^{n} \ell(\widehat{g}_{t-1}, Z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, Z_t) \leqslant \frac{d}{2} \log n + C(\mathcal{F})$$

for $\mathcal{F}$ **bounded** family.

Cumulative risk and online-to-batch conversion

Well-understood **sequential** problem (Merhav 1998, Cesa-Bianchi & Lugosi 2006): there exist $\widehat{g}_0, \dots, \widehat{g}_{n-1}$ s.t., for every $Z_1, \dots, Z_n$,

$$\sum_{t=1}^{n} \ell(\widehat{g}_{t-1}, Z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, Z_t) \leqslant \frac{d}{2} \log n + C(\mathcal{F})$$

for $\mathcal{F}$ **bounded** family. Implies (Barron 1989, Catoni 1997, Yang 2000) **excess risk** of

$$\mathbb{E}[\mathcal{E}(\bar{g}_n)] \leqslant \frac{d \log n}{2n} + \frac{C(\mathcal{F})}{n} \quad \text{for} \quad \bar{g}_n = \frac{1}{n+1} \sum_{t=0}^{n} \widehat{g}_t \,.$$

Cumulative risk and online-to-batch conversion

Well-understood **sequential** problem (Merhav 1998, Cesa-Bianchi & Lugosi 2006): there exist $\widehat{g}_0, \dots, \widehat{g}_{n-1}$ s.t., for every $Z_1, \dots, Z_n$,

$$\sum_{t=1}^{n} \ell(\widehat{g}_{t-1}, Z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, Z_t) \leqslant \frac{d}{2} \log n + C(\mathcal{F})$$

for $\mathcal{F}$ **bounded** family. Implies (Barron 1989, Catoni 1997, Yang 2000) **excess risk** of

$$\mathbb{E}[\mathcal{E}(\bar{g}_n)] \leqslant \frac{d \log n}{2n} + \frac{C(\mathcal{F})}{n} \quad \text{for} \quad \bar{g}_n = \frac{1}{n+1} \sum_{t=0}^{n} \widehat{g}_t \,.$$

- **Distribution-free** bound;
- **Suboptimal rate** for **individual** risk, **inefficient** predictor. **Infinite** for unbounded families, **computational complexity**.

SMP: Sample Minmax Predictor

Conditional density estimation (supervised learning)

- Probabilistic prediction of **label** $y \in \mathcal{Y}$ given **input** $x \in \mathcal{X}$
- **Conditional densities** $f(x) = f(\cdot|x)$ on $\mathcal{Y}$
- **Log-loss** of f at $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ is

$$\ell(f, z) = \ell(f(x), y) = -\log f(y|x)$$

- **Risk** $R(f) = -\mathbb{E}\big[\log f(Y|X)\big]$
- Class $\mathcal{F}$ of conditional densities
- Weak assumptions **on** $P_{Y|X}$: **misspecification** means $p_{Y|X} \notin \mathcal{F}$

SMP: Sample Minmax Predictor

Given **virtual** sample (x, y): **add-one MLE**

$\widehat{f}_n^{x,y} = \operatorname{argmin}_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n \ell(f(X_i), Y_i) + \ell(f(x), y) \right\}$

Sample Minmax Predictor (SMP) is

$$\widetilde{f}_n(x) = \operatorname*{argmin}_g \sup_{y \in \mathcal{Y}} \left\{ \ell(g, y) - \ell(\widehat{f}_n^{x,y}(x), y) \right\}$$

- **Minimizes** general excess risk bound
- "Center" of family of **perturbed MLEs**
- Generally **improper**: $\widetilde{f}_n \notin \mathcal{F}$
- **Regularized** variant

SMP for log-loss

Theorem (M., Gaïffas, 2019)

SMP writes

$$\widetilde{f}_n(y|x) = \frac{\widehat{f}_n^{x,y}(y|x)}{\int_{\mathcal{Y}} \widehat{f}_n^{x,y'}(y'|x)\mu(dy')},$$

and satisfies **excess risk bound**

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}_{Z_1^n,X}\bigg[\log\bigg(\int_{\mathcal{Y}} \widehat{f}_n^{X,y}(y|X)\mu(\mathrm{d}y)\bigg)\bigg].$$

Close to SNML online algorithm (Rissanen–Roos; Kotłowski–Grünwald)

Rhs: **complexity** measure tailored to log-loss (statistical, "localized" analogue of NML-Shtarkov integral from sequential prediction)

Leads to $O(d/n)$ bounds in **misspecified case**

Gaussian linear model

Gaussian linear model

- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, model $\mathcal{F} = \{f_\theta(\cdot|x) = \mathcal{N}(\langle \theta, x \rangle, 1) : \theta \in \mathbb{R}^d\}$

$$\ell(f_\theta, (x, y)) = \frac{1}{2}\big(y - \langle \theta, x \rangle\big)^2 \text{ and } R(f_\theta) = \frac{1}{2}\mathbb{E}\big[(Y - \langle \theta, X \rangle)^2\big]$$

- For **plug-in** estimators $\widehat{f}_n = f_{\widehat{\theta}_n} \in \mathcal{F}$, equivalent to **least-squares regression**
- Otherwise **different problem**: estimating $P_{Y|X}$ vs. $\mathbb{E}[Y|X]$

SMP in the Gaussian linear model

Theorem (SMP for the Gaussian linear model)

SMP is

$$\widetilde{f}_n(\cdot|x) = \mathcal{N}\Big(\langle \widehat{\theta}_n^{\mathrm{LS}}, x\rangle, \big(1 + \langle (n\widehat{\Sigma}_n)^{-1}x, x\rangle\big)^2\Big)$$

and, denoting $\widehat{\ell}_{n+1}$ **leverage score** *of* X_{n+1} *in* X_1^{n+1},

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \underbrace{\mathbb{E}\Big[-\log\big(1-\widehat{\ell}_{n+1}\big)\Big]}_{\text{twice } \textbf{well-specified} \text{ minimax risk}} \underset{\text{(small ball+kurt.)}}{\leqslant} \frac{d}{n}\Big(1 + C\kappa\frac{d}{n}\Big).$$

- Upper bound **does not depend on** $P_{Y|X}$ (only on P_X)
- **Twice well-specified minimax** risk
- $d/n + O((d/n)^2)$ vs $\mathbb{E}[(Y - \langle\theta^*, X\rangle)^2\|\Sigma^{-1/2}X\|^2]/(2n)$ for MLE

Logistic regression

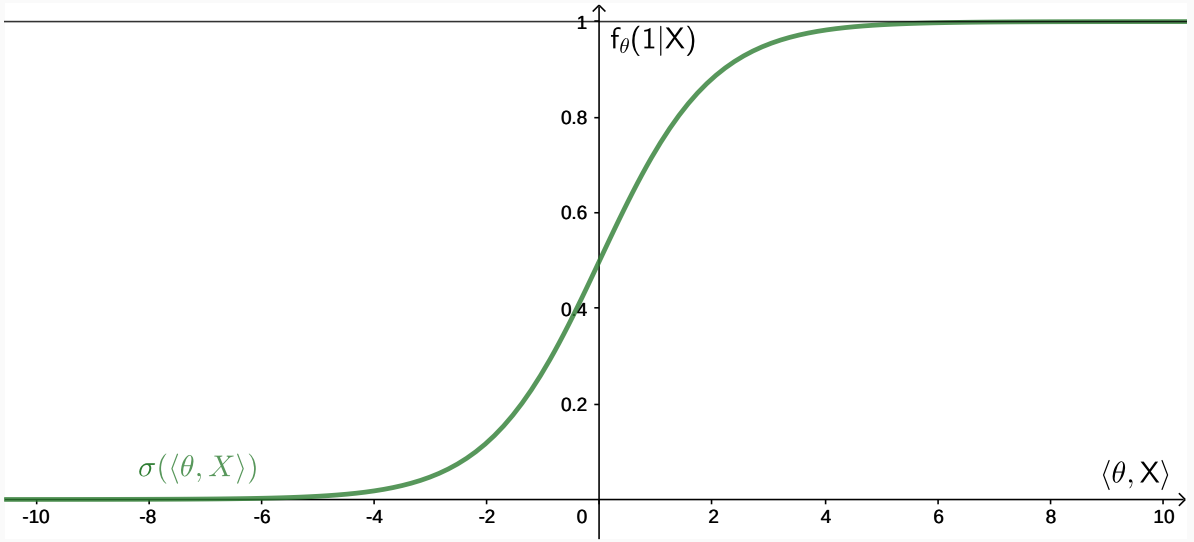
Logistic regression: setting

- **Binary label** $Y \in \{-1, 1\}$  **covariates** $X \in \mathbb{R}^d$. **Risk** of conditional probability $f(\pm 1|x)$

$$R(f) = \mathbb{E}[-\log f(Y|X)]\,.$$

- $\mathcal{F} = \{f_\theta : \theta \in \mathbb{R}^d\}$ **logistic model** of $Y|X$:

$$f_\theta(1|x) = \mathbb{P}_\theta(Y = 1|X = x) = \sigma(\langle\theta, x\rangle), \quad \sigma(u) = \frac{e^u}{1 + e^u} \text{ sigmoid}$$



Learning rates for logistic regression

Assume $\|X\| \leqslant R$, and let $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$. **Excess risk**

$$\mathbb{E}\big[R(f_{\widehat{\theta}_n})\big] - \inf_{f \in \mathcal{F}_B} R(f)\,.$$

Learning rates for logistic regression

Assume $\|X\| \leqslant R$, and let $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$. **Excess risk**

$$\mathbb{E}\big[R(f_{\widehat{\theta}_n})\big] - \inf_{f \in \mathcal{F}_B} R(f)\,.$$

Slow rate of $O(BR/\sqrt{n})$ (convex Lipschitz problem) achieved by:

- **Constrained ERM** over $\mathcal{F}_B$
- **Ridge-ERM** $\widehat{\theta}_{\lambda,n} = \operatorname{argmin}_\theta \widehat{R}(\theta) + \lambda\|\theta\|^2$ with $\lambda = R/(B\sqrt{n})$
- **Projected SGD** on $\mathcal{F}_B$ with averaging, step size $\eta = B/(R\sqrt{n})$

Learning rates for logistic regression

Assume $\|X\| \leqslant R$, and let $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$. **Excess risk**

$$\mathbb{E}\big[R(f_{\widehat{\theta}_n})\big] - \inf_{f \in \mathcal{F}_B} R(f)\,.$$

Slow rate of $O(BR/\sqrt{n})$ (convex Lipschitz problem)

"Fast" rate of $\widetilde{O}(de^{BR}/n)$ (e^{-BR}-exp-concave problem) through:

- **Aggregation with exponential weights** with averaging[2,5], learning rate $\eta = e^{-BR}$
- **Online Newton Step**[5] (averaged)
- **Ridge-ERM**[6,7]

[2]Vovk 1998, [5]Hazan et al. [6]Koren & Levy 2015 [7]Mehta 2017

Learning rates for logistic regression

Assume $\|X\| \leqslant R$, and let $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$. **Excess risk**

$$\mathbb{E}\big[R(f_{\widehat{\theta}_n})\big] - \inf_{f \in \mathcal{F}_B} R(f)\,.$$

Slow rate of $O(BR/\sqrt{n})$ (convex Lipschitz problem)

"Fast" rate of $\widetilde{O}(de^{BR}/n)$ (e^{-BR}-exp-concave problem)

Refined analyses[2] of regularized ERM (self-concordance) $O(\rho d/n)$, ρ distribution-dependent curvature, worst-case $\rho \asymp e^{BR}$

- Asymptotic risk of MLE can be as large as $\sim de^{BR}/n$
- Can be much smaller in practice

[2] Bach 2011, 2014, Bach & Moulines 2013, Ostrovskii & Bach 2018, Marteau-Ferey et al. 2019

Learning rates for logistic regression

Assume $\|X\| \leqslant R$, and let $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$. **Excess risk**

$$\mathbb{E}\big[R(f_{\widehat{\theta}_n})\big] - \inf_{f \in \mathcal{F}_B} R(f)\,.$$

Slow rate of $O(BR/\sqrt{n})$ (convex Lipschitz problem)

"Fast" rate of $\widetilde{O}(de^{BR}/n)$ (e^{-BR}-exp-concave problem)

Refined analyses (self-concordance), $O(de^{BR}/n)$ in worst case

Lower bound[2]: no **proper** estimator $f_{\widehat{\theta}_n} \in \mathcal{F}_B$ can achieve better rate (without further assumptions) than

$$\min\left(\frac{BR}{\sqrt{n}}, \frac{de^{BR}}{n}\right).$$

[2] Hazan et al. 2014

Improper estimators for logistic regression

$\min(de^{BR}/n, BR/\sqrt{n})$ lower bound for **proper** estimators

Does not apply to improper estimators

Online-to-batch conversion of **Bayesian mixture** strategies gives

$$\mathbb{E}R(\overline{f}_n) - \inf_{f \in \mathcal{F}_B} R(f) = O\Big(\frac{d}{n}\log(BRn)\Big).$$

$BR = O(\sqrt{d})$ (natural in finite dim.) leads to $O((d \log n)/n)$ bound

Improper estimators for logistic regression

Averaged Bayesian posteriors: Consider $\overline{f}_n = \frac{1}{n}\sum_{t=1}^{n} \widehat{f}_t$ where

$$\widehat{f}_t(y|x) = \int_{\mathbb{R}^d} \sigma(\langle \theta, x \rangle) \widehat{\pi}_t(d\theta)$$

where $\widehat{\pi}_t$ posterior $\pi(\cdot|X_1, Y_1, \ldots, X_t, Y_t)$ with density (wrt π)

$$\frac{\prod_{s=1}^{t-1} \sigma(-Y_s \langle \theta, X_s \rangle)}{\int_{\mathbb{R}^d} \prod_{s=1}^{t-1} \sigma(-Y_s \langle \theta', X_s \rangle) \pi(d\theta')}$$

Prediction for x **requires approximate posterior sampling**

Open question from Foster et al. (2018). Practical procedure with fast rates?

SMP for logistic regression

SMP-Logistic

- Given $x \in \mathbb{R}^d$, for $y \in \{-1, 1\}$, compute **add-one MLE**

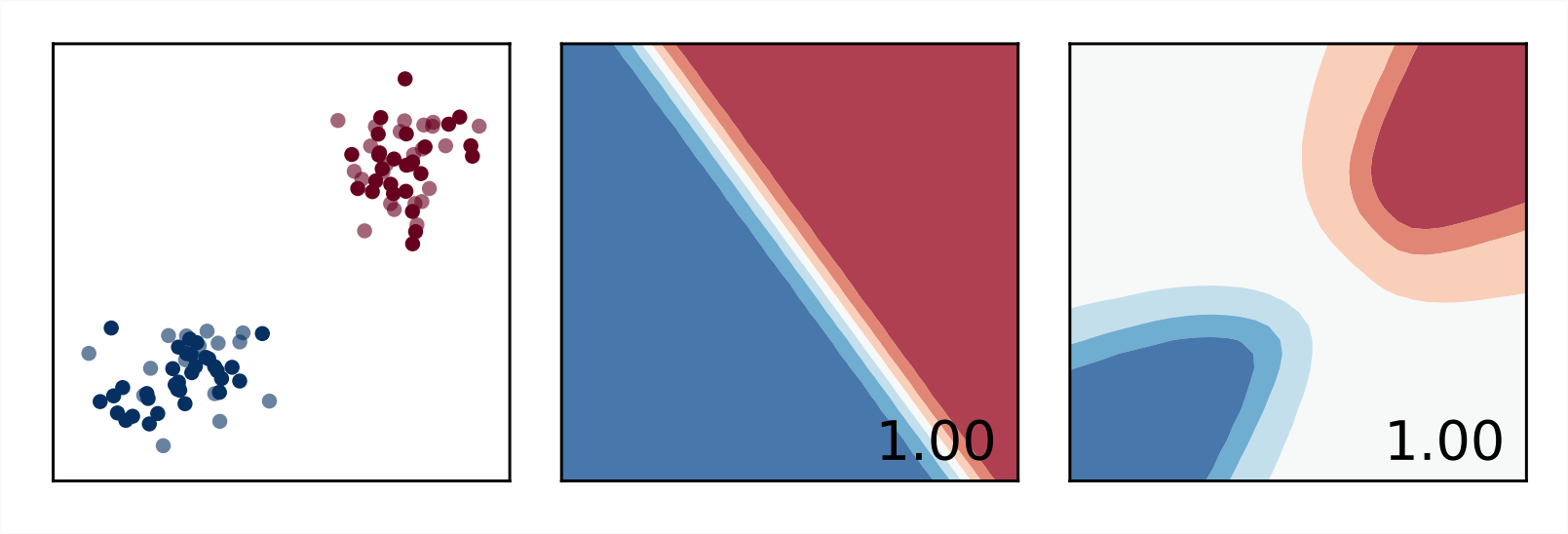
$$\widehat{\theta}_n^{x,y} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(Y_i \langle \theta, X_i \rangle) + \ell(y \langle \theta, x \rangle)$$

- Predict $\mathbb{P}(Y = y | X = x)$ by

$$\widetilde{f}_n(y|x) = \frac{\sigma(y \langle \widehat{\theta}_n^{x,y}, x \rangle)}{\sigma(1 \langle \widehat{\theta}_n^{x,1}, x \rangle) + \sigma(-1 \langle \widehat{\theta}_n^{x,-1}, x \rangle)}$$

- **Replaces posterior sampling by optimization**
- **Non-Bayesian** way to calibrate predictions for logistic regression
- Note: still **more expensive than MLE**: requires to update $\widehat{\theta}_n^{x,\pm 1}$

MLE vs. SMP for logistic regression



Separated data MLE SMP

- MLE **ill-defined**, **over-confident** predictions
- SMP **well-defined**, better **calibrated** predictions ($\widetilde{f}_n(y|x) \in (0, 1)$)

SMP-Ridge for logistic regression

Ridge-MLE augmented by virtual sample (x, y):

$$\widehat{\theta}^{x,y}_{\lambda,n} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\{ \frac{1}{n+1} \Big(\sum_{i=1}^{n} \ell(Y_i \langle \theta, X_i \rangle) + \ell(y \langle \theta, x \rangle) \Big) + \frac{\lambda}{2} \|\theta\|^2 \right\}$$

where $\ell(z) = \log(1 + e^{-z})$. **Ridge-SMP** is

$$\widetilde{f}_{\lambda,n}(y|x) = \frac{\sigma(y\langle \widehat{\theta}^{x,y}_{\lambda,n}, x\rangle) e^{-\lambda \|\widehat{\theta}^{x,y}_{\lambda,n}\|^2/2}}{\sigma(\langle \widehat{\theta}^{x,1}_{\lambda,n}, x\rangle) e^{-\lambda \|\widehat{\theta}^{x,1}_{\lambda,n}\|^2/2} + \sigma(-\langle \widehat{\theta}^{x,-1}_{\lambda,n}, x\rangle) e^{-\lambda \|\widehat{\theta}^{x,-1}_{\lambda,n}\|^2/2}}$$

SMP for logistic regression: guarantees

Theorem (M., Gaïffas, 2019)

Assume $\|X\| \leqslant R$. *Ridge-SMP with* $\lambda = \frac{2R^2}{n+1}$ *satisfies, for any* $B > 0$,

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\|\theta\| \leqslant B} R(f_\theta) \leqslant \frac{3d + B^2R^2}{n}$$

- $O((d + B^2R^2)/n)$; for $BR = O(\sqrt{d})$ gives $O(d/n)$ **bound**
- Bypasses $\min(BR/\sqrt{n}, de^{BR}/n)$ **lower bound** for **proper** estimators (incl. Ridge logistic regression)
- Foster et al. 2018: $O(d \log n/n)$. But more importantly **computationally cheaper**
- **Dimension-free** bound ($\mathrm{tr}[(\Sigma + \lambda I)^{-1}\Sigma]$ instead of d)

Conclusion

MLE (plug-in estimators) overly confident

SMP: procedure for conditional density estimation

Simple with improved guarantees for **logistic** regression

Quantifies **uncertainty** using **virtual sample**, non-Bayesian

Next directions

- **High probability bounds**
- **Online** logistic regression? (recent work by Jézéquel, Gaillard, Rudi)
  Other **online learning** problems?
- Other **generalized linear models**?

Publications

- J.M., O.-A. Maillard. Efficient tracking of a growing number of experts. In *Proc. Algorithmic Learning Theory (ALT)*, 2017
- J.M., S. Gaïffas, E. Scornet. Minimax optimal rates for Mondrian trees and forests. *To appear in Annals of Statistics*, 2020. *NeurIPS 2017*
- J.M., S. Gaïffas, E. Scornet. AMF: Aggregated Mondrian forests for online learning. *In revision*, 2019
- J.M., S. Gaïffas, On the optimality of the Hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 2019
- J. M. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *Submitted*, arXiv:1912.10754, 2019
- J. M., S. Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *Submitted*, arXiv:1912.10784, 2019

Thank you!

Complements

SMP in the Gaussian linear model

Non-uniform bounds

- Before: **uniform** excess risk bounds over $\mathcal{F}$ for $n \gtrsim d$ and P_X non-degenerate (small ball)
- **Non-uniform bounds** over $\mathcal{F}$ relevant when (1) X **not regular**, or (2) $d > n$ (**nonparametric** setting)
- SMP with Ridge penalization $\phi(\theta) = \lambda\|\theta\|^2/2$

Replace previous assumptions by $\|X\| \leqslant R$

- E.g. **bounded kernel** case: $x = \Phi(x')$ with feature map $\Phi : \mathcal{X}' \to \mathbb{R}^d$, such that $K(x', x'') = \langle \Phi(x'), \Phi(x'') \rangle \leqslant R^2$

Remark: parameter scaling (finite dimension)

Assume bounded:

- **Condition number** $c = \|\Sigma\| \cdot \|\Sigma^{-1}\| = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$
- **Leverage** $\|\Sigma^{-1/2}X\| \leqslant \rho\sqrt{d}$ a.s. ($\rho \geqslant 1$ since $\mathbb{E}[\|\Sigma^{-1/2}X\|^2] = d$)
- **Signal strength** $\psi := \|\theta\|_\Sigma = \mathbb{E}[\langle\theta, X\rangle^2]^{1/2}$

Then, if $c, \rho, \psi = O(1)$,

$$\|\theta\| \cdot \|X\| \leqslant c^{1/2}\rho\psi\sqrt{d} \quad \Longrightarrow \quad BR = O(\sqrt{d})$$

Note: we can have $BR \ll \sqrt{d}$ if Σ non-isotropic and θ "aligned" with Σ (non-parametric setting)

SMP in the Gaussian linear model

Degrees of freedom of Ridge estimator (Wahba 1990)

$$\mathrm{df}_\lambda(\Sigma) = \mathrm{tr}[(\Sigma + \lambda I)^{-1}\Sigma]\,.$$

Note that $\mathrm{df}_\lambda(\Sigma) \leqslant d$ and $\mathrm{df}_\lambda(\Sigma) \leqslant R^2/\lambda$ if $\|X\| \leqslant R$.

Theorem (Ridge SMP: nonparametric)

Assume $\mathbb{E}[Y^2] < +\infty$ *and* $\|X\| \leqslant R$. *Then Ridge-SMP satisfies, for* $\lambda \geqslant 2R^2/(n+1)$,

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta\in\mathbb{R}^d}\left\{R(f_\theta) + \frac{\lambda}{2}\|\theta\|^2\right\} \leqslant 1.25\frac{\mathrm{df}_\lambda(\Sigma)}{n+1}$$

- Replaces d by $\mathrm{df}_\lambda(\Sigma)$. **Nonparametric rate**, same bound as well-specified case (minimax over ball), indep. of σ^2, $\mathbb{E}[Y|X]$

SMP in the Gaussian linear model

Proposition (Ridge SMP: finite dimension)

Assume $\mathbb{E}[Y^2] < +\infty$ *and* $\|X\| \leqslant R$. *For any* $B > 0$, *Ridge-SMP* $\widetilde{f}_{\lambda,n}$ *with* $\lambda = \frac{d}{(n+1)B^2}$ *satisfies*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\|\theta\| \leqslant B} R(f_\theta) = O\Big(\frac{d \log(BR/\sqrt{d})}{n}\Big).$$

- If $BR = O(\sqrt{d})$ (finite dimension): $O(d/n)$ bound
- Kakade and Ng (2005): $O(d \log(BRn/\sqrt{d})/n) = O(d \log n/n)$ through sequential problem; SMP removes $\log n$ **factor**
- Also **dimension-free** bounds ($\mathrm{tr}[(\Sigma + \lambda I)^{-1}\Sigma]$ instead of d)

Comparison with stability bound

- **Stability bounds** (Bousquet and Elisseeff 2002) depend on

$$\ell(Y\langle\widehat{\theta}_n^{X,Y}, X\rangle) - \ell(Y\langle\widehat{\theta}_n^{X,-Y}, X\rangle)\ , \qquad \ell(u) = \log(1+e^u)$$

- **SMP bound** depends on $\sigma(Y\langle\widehat{\theta}_n^{X,Y}, X\rangle) - \sigma(Y\langle\widehat{\theta}_n^{X,-Y}, X\rangle)$

When $u \approx u' \gg 1$ we have

$$\ell(u') - \ell(u) \approx u - u'$$
$$\sigma(u') - \sigma(u) \approx e^{-u}(u - u')$$

Explains **why we can remove** e^{BR} factor (next)

SMP for logistic regression: guarantees

Theorem (M., Gaïffas, 2019)

Assume $\|X\| \leqslant R$. *Ridge-SMP satisfies, for* $\lambda \geqslant 2R^2/(n+1)$

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] \leqslant \inf_{\theta \in \mathbb{R}^d} \left\{ R(f_\theta) + \frac{\lambda}{2}\|\theta\|^2 \right\} + \frac{3\,\mathrm{df}_{4\lambda}(\Sigma)}{n}$$

where $\mathrm{df}_\lambda(\Sigma) = \mathrm{tr}\left[(\Sigma + \lambda I)^{-1}\Sigma\right]$.

- Remark: **fast rate** under no assumption on $P_{Y|X}$
- Similar to fast rates in **well-specified case** (Marteau-Ferey et al. 2019) (although more precise bias term in this case)