An improper estimator with optimal excess risk in misspecified density estimation and logistic regression

Jaouad Mourtada*, Stéphane Gaïffas†

StatMathAppli 2019, Fréjus

* CMAP, École polytechnique, † LPSM, Université Paris-Diderot

On arXiv soon.

Predictive density estimation

Predictive density estimation: setting

- Space $\mathcal{Z}$; i.i.d. sample $Z_1^n = (Z_1, \dots, Z_n) \sim P^n$, with P unknown distribution on $\mathcal{Z}$.
- Given Z_1^n, **predict** new sample $Z \sim P$ (probabilistic prediction)
- f density on $\mathcal{Z}$ (wrt base measure μ), $z \in \mathcal{Z}$, **log-loss** $\ell(f, z) = -\log f(z)$. Risk $R(f) = \mathbb{E}[\ell(f, Z)]$ where $Z \sim P$.
- Family $\mathcal{F}$ of densities on $\mathcal{Z}$ = **statistical model**;
- **Goal**: find density $\widehat{g}_n = \widehat{g}_n(Z_1^n)$ with small **excess risk**

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

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

- Standard loss function, connected to lossless compression;
- Minimizing risk amounts to maximizing joint probability attributed to large test sample $(Z'_1, \dots, Z'_m) \sim P^m$:

$$\prod_{j=1}^{m} f(Z'_j) = \exp\left(-\sum_{j=1}^{m} \ell(f, Z'_j) \right) = \exp\left[-m(R(f) + o(1))\right]$$

- Letting $p = dP/d\mu$ be the true density,

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

Risk minimized by **true density**: $f^* = p$; excess risk given by the Kullback-Leibler divergence (relative entropy).

Well-specified case: asymptotic optimality of the MLE

Here, assume that $p \in \mathcal{F}$ (**well-specified** model), with $\mathcal{F}$ a regular parametric family/model of dimension d.

The **Maximum Likelihood Estimator (MLE)** $\widehat{f}_n$, defined by

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

satisfies, as $n \to \infty$,

$$R(\widehat{f}_n) - \inf_{f \in \mathcal{F}} R(f) = \mathrm{KL}(p, \widehat{f}_n) = \frac{d}{2n} + o\left(\frac{1}{n}\right) .$$

The $d/(2n)$ rate is asymptotically optimal (locally asymptotically minimax – Hájek, Le Cam): MLE is **efficient**.

Misspecified case (statistical learning viewpoint)

Assumption $p \in \mathcal{F}$ is restrictive and generally not satisfied: model chosen by the statistician, simplification of the truth.

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

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

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

where $d_{\mathrm{eff}} = \mathrm{Tr}[H^{-1}G]$, $G = \mathbb{E}[\nabla \ell(f^*, Z) \nabla \ell(f^*, Z)^\top]$, $H = \nabla^2 R(f^*)$. Misspecified case: d_{eff} depends on P, and we may have $d_{\mathrm{eff}} \gg d$.

Cumulative risk/regret and online-to-batch conversion

Well-established theory (Merhav 1998, Cesa-Bianchi & Lugosi 2006) for controlling **cumulative** excess risk

$$\mathrm{Regret}_n = \sum_{t=1}^{n} \ell(\widehat{g}_{t-1}, Z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, Z_t)\,;$$

$\mathcal{F}$ **bounded** family: minimax regret of $(d \log n)/2 + O(1)$. Implies excess risk of $(d \log n)/(2n) + O(1/n)$ for **averaged** predictor:

$$\bar{g}_n = \frac{1}{n+1} \sum_{t=0}^{n} \widehat{g}_t\,.$$

$\oplus$ Valid under model misspecification (distribution-free);

$\ominus$ Suboptimal rate for **individual** risk, inefficient predictor. Infinite for unbounded families (eg Gaussian), computational complexity.

The Sample Minimax Predictor

The Sample Minimax Predictor (SMP)

We introduce the **Sample Minimax Predictor**, given by:

$$\widetilde{f}_n = \operatorname*{argmin}_{g} \sup_{z \in \mathcal{Z}} [\ell(g, z) - \ell(\widehat{f}_n^z, z)] = \frac{\widehat{f}_n^z(z)}{\int_{\mathcal{Z}} \widehat{f}_n^{z'}(z')\mu(dz')}$$

where

$$\widehat{f}_n^z = \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \ell(f, Z_i) + \ell(f, z) \right\} .$$

- In general, $\widetilde{f}_n \notin \mathcal{F}$: **improper predictor**.
- **Conditional** variant $\widetilde{f}_n(y|x)$ for conditional density estimation.
- **Regularized** variant.

Excess risk bound for the SMP

$$\widetilde{f}_n(z) = \frac{\widehat{f}_n^{z}(z)}{\int_{\mathcal{Z}} \widehat{f}_n^{z'}(z')\mu(dz')} \tag{1}$$

Theorem (M., Gaïffas, Scornet, 2019)

The SMP $\widetilde{f}_n$ (1) *satisfies:*

$$\mathbb{E}\big[R(\widetilde{f}_n)\big] - \inf_{f \in \mathcal{F}} R(f) \leqslant \mathbb{E}_{Z_1^n}\Big[\log\Big(\int_{\mathcal{Y}} \widehat{f}_n^{(z)}(z)\mu(\mathrm{d}z)\Big)\Big]. \tag{2}$$

- Analogous excess risk bound in the **conditional** case.
- Typically simple $d/n + o(n^{-1})$ bound for standard models (Gaussian, multinomial), even in **misspecified case**.

Application: Gaussian linear model

Gaussian linear model

- Conditional density estimation problem.
- Probabilistic prediction of response $Y \in \mathbf{R}$ given covariates $X \in \mathbf{R}^d$. **Risk** of conditional density $f(y|x)$ is

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

- $\mathcal{F} = \{f_\beta : \beta \in \mathbf{R}^d\}$ with $f_\beta(\cdot|x) = \mathcal{N}(\langle \beta, x\rangle, 1)$, so that

$$\ell(f_\beta, (x, y)) = \frac{1}{2}(y - \langle \beta, x\rangle)^2$$

- MLE is $\widehat{f}_n(\cdot|x) = \mathcal{N}(\langle \widehat{\beta}_n, x\rangle, 1)$, with $\widehat{\beta}_n$ ordinary least squares:

$$\widehat{\beta}_n = \operatorname*{argmin}_{\beta \in \mathbf{R}^d} \sum_{i=1}^n (Y_i - \langle \beta, X_i\rangle)^2 = \left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \sum_{i=1}^n Y_i X_i$$

SMP for the Gaussian linear model

$\Sigma = \mathbb{E}[XX^\top]$, $\widehat{\Sigma}_n = n^{-1} \sum_{i=1}^n X_i X_i^\top$ true/sample covariance matrix

Theorem (Distribution-free excess risk for SMP)

The SMP is $\widetilde{f}_n(\cdot|x) = \mathcal{N}(\langle \widehat{\beta}_n, x\rangle, (1 + \langle (n\widehat{\Sigma}_n)^{-1} x, x\rangle)^2)$. *If* $\mathbb{E}[Y^2] < +\infty$, *then*

$$\mathbb{E}\big[R(\widetilde{f}_n)\big] - \inf_{\beta \in \mathbf{R}^d} R(\beta) \leqslant \mathbb{E}\Big[-\log\Big(1 - \underbrace{\langle (n\widehat{\Sigma}_n + XX^\top)^{-1} X, X\rangle}_{\text{"leverage score"}}\Big)\Big]$$

which is twice the minimax risk in the well-specified case.

- Smaller than $\mathbb{E}[\mathrm{Tr}(\Sigma^{1/2}\widehat{\Sigma}_n^{-1}\Sigma^{1/2})]/n \sim d/n$ under regularity assumption on P_X ($\Sigma^{-1/2}X$ not too close to any hyperplane)
- By contrast, for MLE: $\mathbb{E}[R(\widehat{f}_n)] - R(\beta^*) \sim \mathbb{E}[(Y - \langle \beta^*, X\rangle)^2 \|\Sigma^{-1/2}X\|^2]/(2n)$.

Application to logistic regression

Logistic regression: setting

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

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

- $\mathcal{F} = \{f_\beta : \beta \in \mathbf{R}^d\}$ family of conditional densities of $Y|X$:

$$f_\beta(y|x) = \mathbb{P}_\beta(Y = y|X = x) = \sigma(y\langle \beta, x\rangle)\ , \quad y \in \{-1, 1\}$$

with $\sigma(u) = e^u/(1 + e^u)$ sigmoid function. For $\beta, x \in \mathbf{R}^d$, $y \in \{\pm 1\}$

$$\ell(\beta, (x, y)) = \log(1 + e^{-y\langle \beta, x\rangle})$$

Limitations of MLE and proper (plug-in) predictors

- MLE $f_{\widehat{\beta}_n}(y|x) = \sigma(y\langle \widehat{\beta}_n, x \rangle)$ not fully satisfying for prediction:
  - Ill-defined when sets $\{X_i : Y_i = 1\}$ and $\{X_i : Y_i = -1\}$ are linearly separated, yields 0 or 1 probabilities ($\Rightarrow$ infinite risk).
  - Risk $d_{\text{eff}}/(2n)$; if $\|X\| \leqslant R$, d_{eff} may be as large as[1] $d e^{\|\beta^*\| R}$.
- Lower bound (Hazan et al., 2014) for any **proper** (within class) predictor of $\min(BR/\sqrt{n}, d e^{BR}/n)$.
- Better $O(d \cdot \log(BRn)/n)$ through online-to-batch conversion, with improper predictor (Foster et al., 2018). But computationally expensive (posterior sampling).

[1] Bach & Moulines (2013); see also Ostrovskii & Bach (2018).

Sample Minimax Predictor for logistic regression

The SMP writes:

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

where $\widehat{f}_n^{(x,y)}$ is the MLE obtained when adding (x, y) to the sample.

- Well-defined, even in the separated case; invariant by linear transformation of X ("prior-free"). Never outputs 0 probability.
- Computationally reasonable: prediction obtained by solving two logistic regressions (replaces sampling by optimization).
- NB: still more expensive than simple logistic regression (need to update solution of logistic regression for each test input x).

Excess risk bound for the penalized SMP

Theorem (M., Gaïffas, Scornet 2019)

Assume that $\|X\| \leqslant R$ *a.s. and let* $\lambda = 2R^2/(n+1)$. *Then, logistic SMP with penalty* $\lambda\|\beta\|^2/2$ *satisfies: for every* $\beta \in \mathbf{R}^d$,

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

Remark. Fast rate under no assumption on $\mathcal{L}(Y|X)$.

If $R = O(\sqrt{d})$ and $\|\beta^*\| = O(1)$, then optimal $O(d/n)$ excess risk.

Recall $\min(BR/\sqrt{n}, de^{BR}/n) = \min(\sqrt{d/n}, de^{\sqrt{d}}/n)$ lower bound for proper predictors (incl. Ridge logistic regression).

Also better than $O(d \log n/n)$ from OTB, but worse dependence on $\|\beta^*\|$.

Conclusion

Conclusion

Sample Minimax Predictor = procedure for predictive density estimation. General excess risk bound, typically does not degrade under model misspecification.

Gaussian linear model: tight bound, within a factor of 2 of minimax.

For logistic regression: simple predictor, bypasses lower bounds for proper (plug-in) predictors (removes exponential factor for worst-case distributions).

Next directions:

- Other GLMs?
- Online logistic regression (individual sequences)?
- Application to statistical learning with other loss functions?

Thank you!