Sparse logistic regression: model selection, goodness-of-fit and classification

Felix Abramovich

Tel Aviv University

(based on joint work with Vadim Grinshtein, The Open University of Israel)

Outline

1. Model selection in sparse logistic regression
   - theoretical results: risk bounds, adaptive rate-optimal estimators
   - computational aspects
2. Classification by sparse logistic regression
3. Possible extensions

Sparse logistic regression

- $Y_i \sim B(1, p_i), \quad i = 1, \cdots, n$
- $logit(p_i) = \ln \frac{p_i}{1-p_i} = \beta^t \mathbf{x}_i, \;\; \mathbf{x}_i \in \mathbb{R}^d, \quad logit(\mathbf{p}) = X_{n \times d}\; \beta$

  $rank(X) = r \leq min(n, d)$, any r columns of X are linearly independent

- **Key sparsity assumption**: only some subset of predictors is really "relevant": $||\beta||_0 \leq d_0$

  Goal: to identify this "relevant subset" (the "best" model)

Model selection by penalized MLE

- For a given model $M \subset \{1, \ldots, d\}$,

$$\widehat{\boldsymbol{\beta}}_M = \arg\max_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} \ell(\widetilde{\boldsymbol{\beta}}) = \arg\max_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} \sum_{i=1}^{n} \left\{ \widetilde{\boldsymbol{\beta}}_M^t \, \mathbf{x}_i Y_i - \ln\left(1 + \exp(\widetilde{\boldsymbol{\beta}}_M)^t \mathbf{x}_i\right) \right\}$$

where $\mathcal{B}_M = \{\boldsymbol{\beta} \in \mathbb{R}^d : \beta_j = 0 \text{ if } j \notin M\}$.

$$\widehat{p}_{Mi} = \frac{\exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)}{1 + \exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)}, \quad i = 1, \ldots, n$$

Model selection by penalized MLE

- For a given model $M \subset \{1, \ldots, d\}$,

$$\widehat{\boldsymbol{\beta}}_M = \arg\max_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} \ell(\widetilde{\boldsymbol{\beta}}) = \arg\max_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} \sum_{i=1}^{n} \left\{ \widetilde{\boldsymbol{\beta}}_M^t \, \mathbf{x}_i Y_i - \ln\left(1 + \exp(\widetilde{\boldsymbol{\beta}}_M)^t \mathbf{x}_i\right) \right\}$$

where $\mathcal{B}_M = \{\boldsymbol{\beta} \in \mathbb{R}^d : \beta_j = 0 \text{ if } j \notin M\}$.

$$\widehat{p}_{Mi} = \frac{\exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)}{1 + \exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)}, \quad i = 1, \ldots, n$$

- $\widehat{M} = \arg\min_M \left\{ \sum_{i=1}^n \left(\ln\left(1 + \exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)\right) - \widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i Y_i \right) + Pen(|M|) \right\}$

Model selection by penalized MLE

- For a given model $M \subset \{1, \ldots, d\}$,

$$\widehat{\boldsymbol{\beta}}_M = \arg\max_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} \ell(\widetilde{\boldsymbol{\beta}}) = \arg\max_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} \sum_{i=1}^{n} \left\{ \widetilde{\boldsymbol{\beta}}_M^t \, \mathbf{x}_i Y_i - \ln\left(1 + \exp(\widetilde{\boldsymbol{\beta}}_M)^t \mathbf{x}_i\right) \right\}$$

where $\mathcal{B}_M = \{\boldsymbol{\beta} \in \mathbb{R}^d : \beta_j = 0 \text{ if } j \notin M\}$.

$$\widehat{p}_{Mi} = \frac{\exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)}{1 + \exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)}, \quad i = 1, \ldots, n$$

- $\widehat{M} = \arg\min_M \left\{ \sum_{i=1}^n \left(\ln\left(1 + \exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)\right) - \widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i Y_i \right) + Pen(|M|) \right\}$

Key question: how to choose a "proper" complexity penalty $Pen(|M|)$?

Complexity Penalties

- linear-type penalties $Pen(|M|) = \lambda|M|$

$\lambda = 1$	AIC (Akaike, '73)
$\lambda = \ln(n)/2$	BIC (Schwarz, '78)
$\lambda = \ln d$	RIC (Foster and George, '94)

Complexity Penalties

- linear-type penalties $Pen(|M|) = \lambda|M|$

  $\lambda = 1$ AIC (Akaike, '73)

  $\lambda = \ln(n)/2$ BIC (Schwarz, '78)

  $\lambda = \ln d$ RIC (Foster and George, '94)

- $k\ln(d/k)$-type nonlinear penalties $Pen(|M|) \sim C|M|\ln(de/|M|)$ (Birgé and Massart, '01, '07; Bunea *et al.* '07; AG '10 for Gaussian regression; AG '16 for GLM)

$$k\ln(d/k) \sim \ln\binom{d}{k} \; - \; \log(\text{number of models of size } k)$$

slight modification for $k = r$: $Pen(r) = Cr$

Goodness-of-fit: Kullback-Leibler risk

$$EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) = E\left\{\sum_{i=1}^{n}\left(p_i \ln\left(\frac{p_i}{\widehat{p}_{\widehat{M}i}}\right) + (1-p_i)\ln\left(\frac{1-p_i}{1-\widehat{p}_{\widehat{M}_i}}\right)\right)\right\}$$

Goodness-of-fit: Kullback-Leibler risk

$$EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) = E\left\{\sum_{i=1}^{n}\left(p_i \ln\left(\frac{p_i}{\widehat{p}_{\widehat{M}i}}\right) + (1-p_i)\ln\left(\frac{1-p_i}{1-\widehat{p}_{\widehat{M}_i}}\right)\right)\right\}$$

Assumption (A)

There exists $0 < \delta < 1/2$ *such that* $\delta \le p_i \le 1-\delta$ *or, equivalently,* $|\boldsymbol{\beta}^t \mathbf{x}_i| \le C_0,\ i = 1, \ldots, n$

$(Var(Y_i) = p_i(1-p_i)$ cannot be infinitely close to zero)

Define $\mathcal{B}(d_0) = \{\boldsymbol{\beta} \in \mathbb{R}^d : ||\boldsymbol{\beta}||_0 \le d_0\}$.

Theorem (upper bound, AG '16)

Consider $Pen(k) = Ck \ln\left(\frac{de}{k}\right),\ k = 1, \ldots, r-1$ *and* $Pen(r) = Cr$, *where* $C > \frac{4}{\delta(1-\delta)}$. *Then, under Assumption (A), for some* $C_1 > 0$

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}(d_0)} EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) \le C_1\ \frac{1}{\delta(1-\delta)}\ \min\left(d_0 \ln\left(\frac{de}{d_0}\right), r\right)$$

Theorem (minimax lower bound, AG '16)

Under Assumption (A),

$$\inf_{\widetilde{\mathbf{p}}} \sup_{\boldsymbol{\beta} \in \mathcal{B}(d_0)} EKL(\mathbf{p}, \widetilde{\mathbf{p}}) \ge \begin{cases} C_2\ \delta(1-\delta)\ \tau[2d_0]\ d_0 \ln\left(\frac{de}{d_0}\right), & 1 \le d_0 \le r/2 \\ C_2\ \delta(1-\delta)\ \tau[d_0]\ r, & r/2 \le d_0 \le r \end{cases}$$

where $\tau[k]$ *is the ratio between the minimal and maximal k-sparse eigenvalues of* X

Remarks

- All the above results for model selection in logistic regression can be extended to a general GLM (AG '16):
  - $Y_i \sim f_{\theta_i}(y), \quad f_{\theta_i}(y) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{a} + c(y,a)\right\}$
  - $\theta_i = \boldsymbol{\beta}^t \mathbf{x}_i, \quad \mathbf{x}_i \in \mathbb{R}^d$

  Examples: Gaussian linear regression, logistic regression, Poisson log-linear regression

Remarks

- All the above results for model selection in logistic regression can be extended to a general GLM (AG '16):
  - $Y_i \sim f_{\theta_i}(y), \quad f_{\theta_i}(y) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{a} + c(y,a)\right\}$
  - $\theta_i = \boldsymbol{\beta}^t \mathbf{x}_i, \quad \mathbf{x}_i \in \mathbb{R}^d$

  Examples: Gaussian linear regression, logistic regression, Poisson log-linear regression

- The results can be extended to model selection under additional structural constraints on the set of admissible models (AG '16)

$$Pen(|M|) = C \max(\ln m(|M|), |M|), \text{ where } m(k) = \#\{M : |M| = k\}$$

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}(d_0)} EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) = O\left(\max(\ln m(d_0), d_0)\right)$$

Remarks

- All the above results for model selection in logistic regression can be extended to a general GLM (AG '16):
  - $Y_i \sim f_{\theta_i}(y), \quad f_{\theta_i}(y) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{a} + c(y,a)\right\}$
  - $\theta_i = \boldsymbol{\beta}^t \mathbf{x}_i, \quad \mathbf{x}_i \in \mathbb{R}^d$

  Examples: Gaussian linear regression, logistic regression, Poisson log-linear regression

- The results can be extended to model selection under additional structural constraints on the set of admissible models (AG '16)

$$Pen(|M|) = C \max(\ln m(|M|), |M|), \text{ where } m(k) = \#\{M : |M| = k\}$$

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}(d_0)} EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) = O\left(\max(\ln m(d_0), d_0)\right)$$

- Under Assumption (A), $EKL(\mathbf{p}, \widehat{\mathbf{p}}) \asymp ||\mathbf{X}\beta - \mathbf{X}\widehat{\beta}||^2 \asymp ||\beta - \widehat{\beta}||^2$ all the results remain true (with different constants) for estimating $X\boldsymbol{\beta}$ and $\boldsymbol{\beta}$

Computational aspects

$$\widehat{M} = \arg\min_M \left\{ -\hat{\ell}(M) + Pen(|M|) \right\}$$

combinatorial search over 2^d models (NP problem)

Computational aspects

$$\widehat{M} = \arg\min_{M} \left\{ -\hat{\ell}(M) + Pen(|M|) \right\}$$

combinatorial search over 2^d models (NP problem)

- Greedy algorithms (e.g., forward selection) – approximate the global solution by a stepwise sequence of local ones

(requires strong constraints on X)

Computational aspects

$$\widehat{M} = \arg\min_M \left\{ -\hat{\ell}(M) + Pen(|M|) \right\}$$

combinatorial search over 2^d models (NP problem)

- Greedy algorithms (e.g., forward selection) – approximate the global solution by a stepwise sequence of local ones

  (requires strong constraints on X)

- Convex relaxation methods – replace the original combinatorial problem by some convex surrogate

Convex relaxation methods

(assume that columns of X are normalized to have unit norms)

- Lasso (for linear penalties): $|M| = ||\boldsymbol{\beta}||_0 \to ||\boldsymbol{\beta}||_1$

$$\hat{\boldsymbol{\beta}}_{Lasso} = \arg\min_{\boldsymbol{\beta}} \{-\ell(\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_1\}$$

  - fixed $\lambda \propto \sqrt{\ln d}$
    Under an additional restricted eigenvalue condition on X

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}(d_0)} EKL(\mathbf{p}, \widehat{\mathbf{p}}_{Lasso}) \leq C \, \frac{1}{\delta(1-\delta)} \, d_0 \ln d$$

    (van de Geer '08)
  - adaptively chosen λ (Lepski-type procedure)
    Under somewhat more restrictive conditions on X

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}(d_0)} EKL(\mathbf{p}, \widehat{\mathbf{p}}_{Lasso}) \leq C \, \frac{1}{\delta(1-\delta)} \, d_0 \ln(de/d_0)$$

    (Bellec, Lecué and Tsybakov '16 for Gaussian regression; conjecture for logistic regression)

- SLOPE (Bogdan *et al.* '15): $k\ln(2d/k) \sim \sum_{j=1}^{k} \ln(2d/j)$

$$\hat{\boldsymbol{\beta}}_{Slope} = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j |\beta|_{(j)} \right\}$$

$\lambda_j \propto \sqrt{\ln(2d/j)}$

Under an additional weighted restricted eigenvalue condition on X

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}(d_0)} EKL(\mathbf{p}, \widehat{\mathbf{p}}_{Slope}) \leq C\, \frac{1}{\delta(1-\delta)}\, d_0 \ln(de/d_0)$$

(Su and Candes '16; Bellec, Lecué and Tsybakov '16 for Gaussian regression; AG '17 for GLM)

Classification

- $Y|\mathbf{x} \sim B(1, p(\mathbf{x})), \quad \mathbf{x} \in \mathbb{R}^d$
- Classifier $\eta : \mathbb{R}^d \to \{0, 1\}$
- Missclassification error $R(\eta) = P(Y \neq \eta(\mathbf{x}))$
- Bayes classifier $\eta^*(\mathbf{x}) = I\{p(\mathbf{x}) \geq 1/2\}$

Classification

- $Y|\mathbf{x} \sim B(1, p(\mathbf{x})), \quad \mathbf{x} \in \mathbb{R}^d$
- Classifier $\eta : \mathbb{R}^d \to \{0, 1\}$
- Missclassification error $R(\eta) = P(Y \neq \eta(\mathbf{x}))$
- Bayes classifier $\eta^*(\mathbf{x}) = I\{p(\mathbf{x}) \geq 1/2\}$
- Data $D = ((\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n))$ (fixed or random design)

  (conditional) Missclassification error $R(\hat{\eta}) = P(Y \neq \hat{\eta}(\mathbf{x})|D)$

  Misclassification excess risk $\mathcal{E}(\hat{\eta}, \eta^*) = ER(\hat{\eta}) - R(\eta^*)$

Two main approaches

1. Empirical Risk Minimization (ERM)

$$\hat{\eta} = \arg\min_{\eta \in \mathcal{C}} \hat{R}(\eta) = \arg\min_{\eta \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} I(Y_i \neq \eta(\mathbf{x}_i))$$

- well-developed theory
  (Devroye, Györfi and Lugosi '96; Vapnik '00; see also Boucheron, Bousquet and Lugosi '05 for review)

- computationally infeasible, various convex surrogates (e.g., SVM)

Two main approaches

1. Empirical Risk Minimization (ERM)

$$\hat{\eta} = \arg\min_{\eta \in \mathcal{C}} \hat{R}(\eta) = \arg\min_{\eta \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} I(Y_i \neq \eta(\mathbf{x}_i))$$

   - well-developed theory
     (Devroye, Györfi and Lugosi '96; Vapnik '00; see also Boucheron, Bousquet and Lugosi '05 for review)
   - computationally infeasible, various convex surrogates (e.g., SVM)

2. Plug-in Classifier

   - estimate $p(\mathbf{x})$ from the data
     (e.g, (parametric) logistic regression: $\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^t \mathbf{x}$ or nonparametic – Koltchinskii and Beznosova 05', Audibert and Tsybakov 07')
   - plug-in $\hat{\eta}(\mathbf{x}) = I(\hat{p}(\mathbf{x}) \geq 1/2)$

Logistic regression classifier

1. $\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^t \mathbf{x}$
2. estimate $\boldsymbol{\beta}$ by MLE
3. plug-in $\hat{\eta}(\mathbf{x}) = I(\hat{p}(\mathbf{x}) \geq 1/2) = I(\hat{\boldsymbol{\beta}}^t \mathbf{x} \geq 0)$ – linear classifier

Logistic regression classifier

1. $\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^t \mathbf{x}$

2. estimate $\boldsymbol{\beta}$ by MLE

3. plug-in $\hat{\eta}(\mathbf{x}) = I(\hat{p}(\mathbf{x}) \geq 1/2) = I(\hat{\boldsymbol{\beta}}^t \mathbf{x} \geq 0)$ – linear classifier

For large d classification without feature (model) selection *is as bad as just pure random guessing* (e.g., Bickel and Levina '04; Fan and Fan '08)

Logistic regression classifier

1. $\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^t \mathbf{x}$
2. estimate $\boldsymbol{\beta}$ by MLE
3. plug-in $\hat{\eta}(\mathbf{x}) = I(\hat{p}(\mathbf{x}) \geq 1/2) = I(\hat{\boldsymbol{\beta}}^t \mathbf{x} \geq 0)$ – linear classifier

For large d classification without feature (model) selection *is as bad as just pure random guessing* (e.g., Bickel and Levina '04; Fan and Fan '08)

Sparse logistic regression classifier

1. select the model $\widehat{M}$ (e.g., by penalized MLE – see above)
2. plug-in $\hat{\eta}_{\widehat{M}}(\mathbf{x}) = I(\hat{\boldsymbol{\beta}}^t_{\widehat{M}} \mathbf{x} \geq 0)$

Sparse logistic regression classifier

$$Pen(|M|) = C|M| \ln \frac{de}{|M|}, \quad |M| \le r-1; \quad Pen(r) = C\ r$$

$\mathcal{C}(d_0) = \{\eta(\mathbf{x}) = I\{\boldsymbol{\beta}^t \mathbf{x} \ge 0\} : \boldsymbol{\beta} \in \mathbb{R}^d,\ ||\boldsymbol{\beta}||_0 \le d_0\}$ – the set of d_0-sparse linear classifiers

Lemma (thanks to Noga Alon)

Vapnik-Chervonenkis dimenion $VC(\mathcal{C}(d_0)) \sim d_0 \ln\left(\frac{de}{d_0}\right)$ *:*

$$d_0 \log_2\left(\frac{2d}{d_0}\right) \le VC(\mathcal{C}(d_0)) \le 2d_0 \log_2\left(\frac{de}{d_0}\right)$$

Hence, $Pen(|M|) \propto VC(\mathcal{C}(|M|))$

Fixed design; average misclassification error $R_X(\eta) = \frac{1}{n}\sum_{i=1}^{n} P(Y_i \neq \eta(\mathbf{x}_i))$

$$\mathcal{E}_X(\hat{\eta}, \eta^*) = ER_X(\hat{\eta}) - R_X(\eta^*)$$

- Assumption (A): $\delta \leq p_i \leq 1-\delta, \quad i = 1, \ldots, n$

Fixed design; average misclassification error $R_X(\eta) = \frac{1}{n}\sum_{i=1}^n P(Y_i \neq \eta(\mathbf{x}_i))$

$$\mathcal{E}_X(\hat{\eta}, \eta^*) = ER_X(\hat{\eta}) - R_X(\eta^*)$$

- Assumption (A): $\delta \leq p_i \leq 1-\delta, \quad i = 1, \ldots, n$
- $\sup_{\boldsymbol{\beta} \in \mathcal{B}(d_0)} EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) \leq O\left(\min\left(d_0 \ln \frac{de}{d_0}, r\right)\right)$

Fixed design; average misclassification error $R_X(\eta) = \frac{1}{n}\sum_{i=1}^n P(Y_i \neq \eta(\mathbf{x}_i))$

$$\mathcal{E}_X(\hat{\eta}, \eta^*) = ER_X(\hat{\eta}) - R_X(\eta^*)$$

- Assumption (A): $\delta \leq p_i \leq 1 - \delta, \quad i = 1, \ldots, n$
- $\sup_{\boldsymbol{\beta} \in \mathcal{B}(d_0)} EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) \leq O\left(\min\left(d_0 \ln \frac{de}{d_0}, r\right)\right)$

  (see above)
- $\mathcal{E}_X(\hat{\eta}_{\widehat{M}}, \eta^*) \leq \sqrt{\frac{2}{n}\ EKL(\mathbf{p}, \hat{\mathbf{p}}_{\widehat{M}})}$

  (Zhang '04; Bartlett, Jordan and McAuliffe '06)

Excess risk bounds

Theorem (upper bound)

Under Assumption (A), for the $k\ln(d/k)$*-type complexity penalty,*

$$\sup_{\eta \in \mathcal{C}(d_0)} \mathcal{E}_X(\hat{\eta}_{\widehat{M}}, \eta^*) \leq C_1 \sqrt{\frac{1}{\delta(1-\delta)} \; \frac{\min\left(d_0 \ln \frac{de}{d_0}, r\right)}{n}}$$

Excess risk bounds

Theorem (upper bound)

Under Assumption (A), for the $k\ln(d/k)$*-type complexity penalty,*

$$\sup_{\eta\in\mathcal{C}(d_0)} \mathcal{E}_X(\hat{\eta}_{\widehat{M}}, \eta^*) \le C_1 \sqrt{\frac{1}{\delta(1-\delta)}\ \frac{\min\left(d_0 \ln \frac{de}{d_0}, r\right)}{n}}$$

Theorem (minimax lower bound)

Consider a sparse agnostic $(R(\eta^* > 0))$ *logistic regression model with* $2 \le d_0 \ln \frac{2d}{d_0} \le n$*. There exists a design matrix* X_0 *such that*

$$\inf_{\tilde{\eta}} \sup_{\eta*\in\mathcal{C}(d_0)} \mathcal{E}_{X_0}(\tilde{\eta}, \eta^*) \ge C_2 \sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}}$$

or (see Lemma) $\inf_{\tilde{\eta}} \sup_{\eta*\in\mathcal{C}(d_0)} \mathcal{E}_{X_0}(\tilde{\eta}, \eta^*) \ge C_2 \sqrt{\frac{VC(\mathcal{C}(d_0))}{n}}$

Extensions. Tighter risk bounds under low-noise condition

Assume an additional low-noise condition (p_i are separated from $1/2$):

$$h \leq |p_i - 1/2| \leq \Delta, \quad i = 1, \ldots, n$$

(Massart and Nédélec '06)

$$\mathcal{E}_X(\widehat{\eta}_{\widehat{M}}, \eta^*) = O\left(\min\left\{\sqrt{\frac{\min\left(d_0 \ln \frac{de}{d_0}, r\right)}{n}}, \frac{\min\left(d_0 \ln \frac{de}{d_0}, r\right)}{nh}\right\}\right)$$

- the excess risk is reduced for $h > \sqrt{\frac{\min\left(d_0 \ln \frac{de}{d_0}, r\right)}{n}}$
- $\widehat{\eta}_{\widehat{M}}$ is rate-optimal classifier (in terms of "the worst case design")

Random design (in progress)

$$(\mathbf{X}, Y) \sim \mathcal{F} : Y|\mathbf{x} \sim B(1, p(\mathbf{x})),\ \ln \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = \boldsymbol{\beta}^t \mathbf{x},\ ||\boldsymbol{\beta}||_0 \leq d_0;\ \mathbf{X} \sim f(\mathbf{x})$$

- minimax lower bound (Devroye, Györfi and Lugosi, '96 + Lemma)

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}(d_0), h} \mathcal{E}(\tilde{\eta}, \eta^*) \geq \sqrt{\frac{V(\mathcal{C}(d_0))}{n}} \sim C \sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}}$$

- Assume
  $\delta \leq p(\mathbf{x}) \leq 1 - \delta,\ \ 0 < \delta < 1/2; \quad f(\mathbf{x}) \geq \gamma > 0,\ \mathbf{x} \in supp\ f(\mathbf{x})$

- upper bound for $k \ln(d/k)$-type complexity penalty

$$\sup_{\eta^* \in \mathcal{C}(d_0)} \mathcal{E}(\hat{\eta}_{\widehat{M}}, \eta^*) = O\left(\sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}}\right)$$

- The rates can be improved under additional low-noise conditions

Multiclass classification by multinomial logistic regression (in progress)

$$\mathbf{Y} \sim Multinom(p_1(\mathbf{x}), \ldots, p_L(\mathbf{x})), \quad \mathbf{Y} \in \{0,1\}^L, \quad \mathbf{x} \in \mathbb{R}^d$$
$$\sum_{j=1}^L Y_j = 1, \; \sum_{j=1}^L p_j(\mathbf{x}) = 1$$

$$\theta_j = \ln \frac{p_j(\mathbf{x})}{p_L(\mathbf{x})} = \boldsymbol{\beta}_j^t \mathbf{x}, \quad p_j(\mathbf{x}) = \frac{\exp\left(\boldsymbol{\beta}_j^t \mathbf{x}\right)}{\sum_{k=1}^L \exp\left(\boldsymbol{\beta}_k^t \mathbf{x}\right)}, \quad \boldsymbol{\beta}_L = \mathbf{0}, \quad j = 1, \ldots, L$$

Multiclass classification by multinomial logistic regression (in progress)

$$\mathbf{Y} \sim Multinom(p_1(\mathbf{x}), \ldots, p_L(\mathbf{x})), \quad \mathbf{Y} \in \{0,1\}^L, \quad \mathbf{x} \in \mathbb{R}^d$$
$$\sum_{j=1}^L Y_j = 1, \ \sum_{j=1}^L p_j(\mathbf{x}) = 1$$

$$\theta_j = \ln \frac{p_j(\mathbf{x})}{p_L(\mathbf{x})} = \boldsymbol{\beta}_j^t \mathbf{x}, \quad p_j(\mathbf{x}) = \frac{\exp\left(\boldsymbol{\beta}_j^t \mathbf{x}\right)}{\sum_{k=1}^L \exp\left(\boldsymbol{\beta}_k^t \mathbf{x}\right)}, \quad \boldsymbol{\beta}_L = \mathbf{0}, \quad j = 1, \ldots, L$$

- Classifier $\eta : \mathbb{R}^d \to \{1, \ldots, L\}$
- Bayes classifier $\eta^*(\mathbf{x}) = \arg\max_{1 \le j \le L} \ p_j(\mathbf{x})$

Sparse multinomial logistic regression

$$\mathbf{Y}_i \sim Multinom(p_1(\mathbf{x}_i), \ldots, p_L(\mathbf{x}_i)), \quad i = 1, \ldots, n$$

$$\Theta_{n\times(L-1)} = X_{n\times d}\ B_{d\times(L-1)}$$

- For a given model $M \subset \{1, \ldots, d\}$, $|M| = \#\{\mathrm{non-zero\ rows}(B)\}$
- $$\widehat{M} = \arg\min_M \left\{ -\hat{\ell}(M) + C(L-1)|M| \ln \frac{de}{|M|} \right\}$$
- Assumption (A'): $0 < \delta \le p_{(1)}(\mathbf{x}) \le \ldots \le p_{(L)}(\mathbf{x}) \le 1-\delta$
- Under Assumption (A')

$$\sup_{B\in\mathcal{B}(d_0)} EKL(P, \widehat{P}_{\widehat{M}}) \le C_1\ \frac{L-1}{\delta}\ \min\left(d_0 \ln\left(\frac{de}{d_0}\right), r\right),$$

where $\mathcal{B}_{d_0} = \{B \in \mathbb{R}^{d\times(L-1)} : \#\{non-zero\ rows(B)\} \le d_0\}$

Sparse multinomial logistic classifier

$$\hat{\eta}_{\widehat{M}}(\mathbf{x}) = \arg\max_{1 \le j \le L} \hat{p}_{\widehat{M}j}(\mathbf{x}) = \arg\max_{1 \le j \le L} \widehat{\beta}^t_{\widehat{M}j}\mathbf{x}$$

Under Assumption (A')

$$\sup_{\eta \in \mathcal{C}(d_0)} \mathcal{E}_X(\hat{\eta}_{\widehat{M}}, \eta^*) = O\left(\sqrt{\frac{L-1}{\delta}\;\frac{min(d_0 \ln \frac{de}{d_0}, r)}{n}}\right)$$

Thank You!