

High-dimensional classification by sparse logistic regression

Felix Abramovich

Tel Aviv University

(joint work with Vadim Grinshtein, The Open University of Israel and Tomer Levy, Tel Aviv University)

Outline

- 1 Review on (binary) classification
- 2 High-dimensional (binary) classification by sparse logistic regression
 - ▶ model, feature selection by penalized maximum likelihood
 - ▶ theory: misclassification excess bounds, adaptive minimax classifiers
 - ▶ computational issues: logistic Lasso and Slope
- 3 Multiclass extensions
 - ▶ sparse multinomial logistic regression
 - ▶ theory: misclassification excess bounds, phase transition
 - ▶ multinomial logistic group Lasso and Slope

Binary Classification

- $(\mathbf{X}, Y) \sim \mathcal{F} : Y|\mathbf{X} = \mathbf{x} \sim B(1, p(\mathbf{x})), \mathbf{X} \in \mathbb{R}^d \sim f(\mathbf{x})$
- Classifier $\eta : \mathbb{R}^d \rightarrow \{0, 1\}$
- Missclassification error $R(\eta) = P(Y \neq \eta(\mathbf{x}))$
- Bayes classifier $\eta^*(\mathbf{x}) = \arg \min_{\eta} R(\eta)$

$$\eta^*(\mathbf{x}) = I\{p(\mathbf{x}) \geq 1/2\}, \quad R(\eta^*) = E_{\mathbf{X}} (\min(p(\mathbf{X}), 1 - p(\mathbf{X})))$$

Binary Classification

- $(\mathbf{X}, Y) \sim \mathcal{F} : Y|\mathbf{X} = \mathbf{x} \sim B(1, p(\mathbf{x})), \mathbf{X} \in \mathbb{R}^d \sim f(\mathbf{x})$
- Classifier $\eta : \mathbb{R}^d \rightarrow \{0, 1\}$
- Missclassification error $R(\eta) = P(Y \neq \eta(\mathbf{x}))$
- Bayes classifier $\eta^*(\mathbf{x}) = \arg \min_{\eta} R(\eta)$

$$\eta^*(\mathbf{x}) = I\{p(\mathbf{x}) \geq 1/2\}, \quad R(\eta^*) = E_{\mathbf{X}}(\min(p(\mathbf{X}), 1 - p(\mathbf{X})))$$

- Data $D = (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n) \sim \mathcal{F}$

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

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

Vapnik-Chervonenkis (VC) dimension

Definition

Let \mathcal{C} be a set of classifiers. $VC(\mathcal{C})$ is the maximal number of points in \mathcal{X} that can be arbitrarily classified by classifiers in \mathcal{C} .

Vapnik-Chervonenkis (VC) dimension

Definition

Let \mathcal{C} be a set of classifiers. $VC(\mathcal{C})$ is the maximal number of points in \mathcal{X} that can be arbitrarily classified by classifiers in \mathcal{C} .

Example: linear classifiers $\mathcal{C}_L = \{\eta(\mathbf{x}) = I\{\beta^T \mathbf{x} \geq 0\}, \beta \in \mathbb{R}^d\}$, $VC(\mathcal{C}_L) = d$

Vapnik-Chervonenkis (VC) dimension

Definition

Let \mathcal{C} be a set of classifiers. $VC(\mathcal{C})$ is the maximal number of points in \mathcal{X} that can be arbitrarily classified by classifiers in \mathcal{C} .

Example: linear classifiers $\mathcal{C}_L = \{\eta(\mathbf{x}) = I\{\beta^T \mathbf{x} \geq 0\}, \beta \in \mathbb{R}^d\}$, $VC(\mathcal{C}_L) = d$

Minimax lower bound

Let $2 \leq VC(\mathcal{C}) < \infty$, $n \geq VC(\mathcal{C})$ and $R(\eta^*) > 0$.

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}, f(\mathbf{x})} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C \sqrt{\frac{VC(\mathcal{C})}{n}}$$

(e.g., Devroye, Györfi and Lugosi, '96)

In particular, for linear classifiers $\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_L, f(\mathbf{x})} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C \sqrt{\frac{d}{n}}$

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)

$$\sup_{\eta^* \in \mathcal{C}} \mathcal{E}(\hat{\eta}, \eta^*) \leq C \sqrt{\frac{VC(\mathcal{C})}{n}} \quad (\text{optimal order})$$

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

2. Plug-in Classifiers

- estimate $p(\mathbf{x})$ from the data
(e.g, (parametric) logistic regression: $\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \beta^T \mathbf{x}$ or
nonparametric: Yang '99, Koltchinskii and Beznosova '05, Audibert
and Tsybakov '07)
- plug-in $\hat{\eta}(\mathbf{x}) = I(\hat{p}(\mathbf{x}) \geq 1/2)$

2. Plug-in Classifiers

- estimate $p(\mathbf{x})$ from the data
(e.g, (parametric) logistic regression: $\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \beta^T \mathbf{x}$ or
nonparametric: Yang '99, 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 $\ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \beta^T \mathbf{x}$
- 2 estimate β by MLE
- 3 plug-in $\hat{\eta} = I(\hat{p}(\mathbf{x}) \geq 1/2) = I(\hat{\beta}^T \mathbf{x} \geq 0)$ – linear classifier

Big Data era – curse of dimensionality

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)

Big Data era – curse of dimensionality

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 model/feature selection – \hat{M}
- 2 plug-in $\hat{\eta}_{\hat{M}} = I(\hat{\beta}_{\hat{M}}^T \mathbf{x} \geq 0)$

Sparse logistic regression

- $(\mathbf{X}, Y) \sim \mathcal{F} : Y|\mathbf{X} = \mathbf{x} \sim B(1, p(\mathbf{x})), \mathbf{X} \in \mathbb{R}^d \sim f(\mathbf{x})$
- $\text{logit}(p(\mathbf{x})) = \ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^T \mathbf{x}$
- sparsity assumption: $\|\boldsymbol{\beta}\|_0 \leq d_0$

Sparse logistic regression

- $(\mathbf{X}, Y) \sim \mathcal{F} : Y | \mathbf{X} = \mathbf{x} \sim B(1, p(\mathbf{x})), \mathbf{X} \in \mathbb{R}^d \sim f(\mathbf{x})$
- $\text{logit}(p(\mathbf{x})) = \ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \boldsymbol{\beta}^T \mathbf{x}$
- sparsity assumption: $\|\boldsymbol{\beta}\|_0 \leq d_0$

Lemma (thanks to Noga Alon)

Let $\mathcal{C}_L(d_0) = \{\eta(\mathbf{x}) = I\{\boldsymbol{\beta}^T \mathbf{x} \geq 0\} : \boldsymbol{\beta} \in \mathbb{R}^d, \|\boldsymbol{\beta}\|_0 \leq d_0\}$.

$$d_0 \log_2 \left(\frac{2d}{d_0} \right) \leq VC(\mathcal{C}_L(d_0)) \leq 2d_0 \log_2 \left(\frac{de}{d_0} \right), \text{ i.e.}$$

$$VC(\mathcal{C}_L(d_0)) \sim d_0 \ln \left(\frac{de}{d_0} \right)$$

Model/feature selection by penalized MLE

- For a given model $M \subseteq \{1, \dots, d\}$,

$$\hat{\beta}_M = \arg \max_{\tilde{\beta} \in \mathcal{B}_M} \sum_{i=1}^n \left\{ \tilde{\beta}_M^T \mathbf{x}_i Y_i - \ln \left(1 + \exp(\tilde{\beta}_M^T \mathbf{x}_i) \right) \right\},$$

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

Model/feature selection by penalized MLE

- For a given model $M \subseteq \{1, \dots, d\}$,

$$\hat{\beta}_M = \arg \max_{\tilde{\beta} \in \mathcal{B}_M} \sum_{i=1}^n \left\{ \tilde{\beta}_M^T \mathbf{x}_i Y_i - \ln \left(1 + \exp(\tilde{\beta}_M^T \mathbf{x}_i) \right) \right\},$$

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

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

$$\hat{p}_{\hat{M}}(\mathbf{x}) = \frac{\exp(\hat{\beta}_{\hat{M}}^T \mathbf{x})}{1 + \exp(\hat{\beta}_{\hat{M}}^T \mathbf{x})}$$

$$\hat{\eta}_{\hat{M}}(\mathbf{x}) = I(\hat{p}_{\hat{M}}(\mathbf{x}) \geq 1/2) = I(\hat{\beta}_{\hat{M}}^T \mathbf{x} \geq 0)$$

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 $Pen(|M|) = \lambda|M|$

$$\lambda = 1 \quad \text{AIC (Akaike, '73)}$$

$$\lambda = \ln(n)/2 \quad \text{BIC (Schwarz, '78)}$$

$$\lambda = \ln d \quad \text{RIC (Foster and George, '94)}$$

- $k \ln(d/k)$ -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)

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

In addition, for classification, $k \ln(d/k) \sim VC(\mathcal{C}_L(k))$ (recall Lemma)

Assumption (boundedness)

There exists $0 < \delta < 1/2$ such that $\delta < p(\mathbf{x}) < 1 - \delta$ or, equivalently, there exists $C_0 > 0$ such that $|\beta^T \mathbf{x}| < C_0$ for all $\mathbf{x} \in \mathcal{X}$.

The assumption prevents the variance $\text{Var}(Y) = p(\mathbf{x})(1 - p(\mathbf{x}))$ to be arbitrarily close to zero.

Excess risk bounds

Theorem (upper bound)

Under the boundedness assumption, for $\text{Pen}(|M|) = C|M| \ln\left(\frac{de}{|M|}\right)$,

$$\sup_{\eta \in \mathcal{C}_L(d_0)} \mathcal{E}(\hat{\eta}_{\hat{M}}, \eta^*) \leq C(\delta) \sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}}$$

The idea of the proof:

- 1 $\mathcal{E}(\hat{\eta}_{\hat{M}}, \eta^*) \leq \sqrt{2 \text{EKL}(p^*, \hat{p}_{\hat{M}})}$ (Zhang '04; Bartlett *et al.* '06)
- 2 $\sup_{\beta \in \mathcal{B}(d_0)} \text{EKL}(p^*, \hat{p}_{\hat{M}}) = O\left(\frac{d_0 \ln \frac{de}{d_0}}{n}\right)$ (AG '16)

Excess risk bounds

Theorem (upper bound)

Under the boundedness assumption, for $\text{Pen}(|M|) = C|M| \ln\left(\frac{de}{|M|}\right)$,

$$\sup_{\eta \in \mathcal{C}_L(d_0)} \mathcal{E}(\hat{\eta}_{\hat{M}}, \eta^*) \leq C(\delta) \sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}}$$

The idea of the proof:

- 1 $\mathcal{E}(\hat{\eta}_{\hat{M}}, \eta^*) \leq \sqrt{2 \text{EKL}(p^*, \hat{p}_{\hat{M}})}$ (Zhang '04; Bartlett *et al.* '06)
- 2 $\sup_{\beta \in \mathcal{B}(d_0)} \text{EKL}(p^*, \hat{p}_{\hat{M}}) = O\left(\frac{d_0 \ln \frac{de}{d_0}}{n}\right)$ (AG '16)

Recall the **lower bound** for $2 \leq d_0 \ln\left(\frac{de}{d_0}\right) \leq n$:

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_L(d_0), f(\mathbf{x})} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C \sqrt{\frac{VC(\mathcal{C}_L(d_0))}{n}} \geq C \sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}}$$

Tighter bounds under the additional low-noise condition

The main challenges are near the hyperplane $\beta^T \mathbf{x} = 0$, where $p(\mathbf{x}) = 1/2$.

Assumption (low-noise condition)

$P(|p(\mathbf{X}) - 1/2| \leq h) \leq Ch^\alpha, \alpha \geq 0$ for all $0 < h < h^*$ (Tsybakov '04)

- $\alpha = 0$ – no assumptions on the noise (as previously)
- $\alpha = \infty$ – there exists the “corridor” of width $2 \ln \frac{1+2h^*}{1-2h^*}$ that separates the sets $\{\mathbf{x} : \beta^T \mathbf{x} > 0\}$ and $\{\mathbf{x} : \beta^T \mathbf{x} < 0\}$

Tighter bounds under the additional low-noise condition

The main challenges are near the hyperplane $\beta^T \mathbf{x} = 0$, where $p(\mathbf{x}) = 1/2$.

Assumption (low-noise condition)

$$P(|p(\mathbf{X}) - 1/2| \leq h) \leq Ch^\alpha, \alpha \geq 0 \text{ for all } 0 < h < h^* \text{ (Tsybakov '04)}$$

- $\alpha = 0$ – no assumptions on the noise (as previously)
- $\alpha = \infty$ – there exists the “corridor” of width $2 \ln \frac{1+2h^*}{1-2h^*}$ that separates the sets $\{\mathbf{x} : \beta^T \mathbf{x} > 0\}$ and $\{\mathbf{x} : \beta^T \mathbf{x} < 0\}$

Under the low-noise assumption, for all $1 \leq d_0 \leq \min(d, n)$ and all $\alpha \geq 0$,

$$\sup_{\eta \in \mathcal{C}_L(d_0)} \mathcal{E}(\hat{\eta}_{\hat{M}}, \eta^*) \leq \left(C(\delta) \frac{d_0 \ln \frac{de}{d_0}}{n} \right)^{\frac{\alpha+1}{\alpha+2}}$$

$\hat{\eta}_{\hat{M}}$ is rate-optimal and adaptive to both d_0 and α .

Computational aspects

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

combinatorial search over 2^d models (NP problem)

Computational aspects

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

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
(require strong constraints on design)

Computational aspects

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

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
(require strong constraints on design)
- **Convex relaxation methods** – replace the original **combinatorial** problem by some **convex** surrogate

Assumption on design

Assumption (design)

- 1 $EX_j^2 = 1$ (X_j 's are scaled)
- 2 $E(|X_j|^p)^{1/p} \lesssim \sqrt{p}$ for all $2 \leq p \lesssim \ln d$

(e.g., Gaussian, sub-Gaussian)

Assumption ensures that for $n \gtrsim \ln d$,

$$E \max_{1 \leq j \leq d} \frac{1}{n} \sum_{i=1}^n X_{ij}^2 \leq C$$

(Lecué and Mendelson '17)

Convex relaxation methods

Assume boundedness and design assumptions, $n \gtrsim \ln d$

- **logistic Lasso** (for **linear** penalties): $\|\hat{\beta}\|_0 \rightarrow \|\hat{\beta}\|_1$

$$\hat{\beta}_{Lasso} = \arg \min_{\beta} \left\{ -\frac{1}{n} \ell(\beta) + \lambda \|\beta\|_1 \right\}$$

- ▶ **fixed** $\lambda \propto \sqrt{\frac{\ln d}{n}}$: **rate-suboptimal** (up to an extra log-factor:

$$O\left(\sqrt{\frac{d_0 \ln d}{n}}\right) \quad (\text{van de Geer '08, Bellec et al. '16})$$

- ▶ **adaptively chosen** λ (Lepski procedure): **rate-optimal** ($O\left(\sqrt{\frac{d_0 \ln(d_e/d_0)}{n}}\right)$) (Bellec et al. '16 for Gaussian regression; conjecture for classification)

Convex relaxation methods

Assume boundedness and design assumptions, $n \gtrsim \ln d$

- **logistic Lasso** (for **linear** penalties): $\|\hat{\beta}\|_0 \rightarrow \|\hat{\beta}\|_1$

$$\hat{\beta}_{Lasso} = \arg \min_{\beta} \left\{ -\frac{1}{n} \ell(\beta) + \lambda \|\beta\|_1 \right\}$$

- ▶ **fixed** $\lambda \propto \sqrt{\frac{\ln d}{n}}$: **rate-suboptimal** (up to an extra log-factor:

$$O\left(\sqrt{\frac{d_0 \ln d}{n}}\right) \quad (\text{van de Geer '08, Bellec et al. '16})$$

- ▶ **adaptively chosen** λ (Lepski procedure): **rate-optimal** ($O\left(\sqrt{\frac{d_0 \ln(de/d_0)}{n}}\right)$) (Bellec et al. '16 for Gaussian regression; conjecture for classification)

- **logistic Slope**: $k \ln(2d/k) \sim \sum_{j=1}^k \ln(2d/j)$

$$\hat{\beta}_{Slope} = \arg \min_{\beta} \left\{ -\frac{1}{n} \ell(\beta) + \sum_{j=1}^d \lambda_j |\beta|_{(j)} \right\}, \quad \lambda_1 \geq \dots \geq \lambda_d > 0$$

$$\lambda_j \propto \sqrt{\frac{\ln(2d/j)}{n}} : \text{rate-optimal} \left(O\left(\sqrt{\frac{d_0 \ln(de/d_0)}{n}}\right) \right) \quad (\text{AG '19})$$

Misspecified logistic model

$$R(\hat{\eta}_{\hat{M}}) - R(\eta^*) = \underbrace{(R(\hat{\eta}_{\hat{M}}) - R(\eta_L^*))}_{\text{estimation error}} + \underbrace{(R(\eta_L^*) - R(\eta^*))}_{\text{approximation error}}$$

$\eta_L^* = \arg \min_{\eta \in \mathcal{C}_L(d)} R(\eta)$ is the best (oracle) linear classifier

Multiclass classification

- appears in a variety applications, a lot of methods
- much less theory behind

Multiclass classification

- appears in a variety applications, a lot of methods
- much less theory behind

Main approaches :

- ➊ reduction to a series of **binary** classifications
 - ▶ **One-vs-All** – each class is compared against all others
 - ▶ **One-vs-One** – all pairs of classes are compared to each other
- ➋ extensions of binary classification approaches

Multiclass classification

- $(\mathbf{X}, Y) \sim \mathcal{F} : Y | \mathbf{X} = \mathbf{x} \sim \text{Mult}(p_1(\mathbf{x}), \dots, p_L(\mathbf{x}))$, $\mathbf{X} \in \mathbb{R}^d \sim f(\mathbf{x})$
- Classifier $\eta : \mathbb{R}^d \rightarrow \{1, \dots, L\}$
- Missclassification error $R(\eta) = P(Y \neq \eta(\mathbf{x}))$
- Bayes classifier $\eta^*(\mathbf{x}) = \arg \max_{1 \leq j \leq L} p_j(\mathbf{x})$,
 $R(\eta^*) = 1 - E_{\mathbf{X}}(\max_{1 \leq j \leq L} p_j(\mathbf{X}))$

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

Multinomial logistic regression

$$Y|\mathbf{X} = \mathbf{x} \sim \text{Mult}(p_1(\mathbf{x}), \dots, p_L(\mathbf{x})), \quad \mathbf{x} \in \mathbb{R}^d, \quad \sum_{j=1}^L p_j(\mathbf{x}) = 1$$

$$\theta_j = \ln \frac{p_j(\mathbf{x})}{p_L(\mathbf{x})} = \beta_j^T \mathbf{x}, \quad p_j(\mathbf{x}) = \frac{\exp(\beta_j^T \mathbf{x})}{\sum_{k=1}^L \exp(\beta_k^T \mathbf{x})}, \quad j = 1, \dots, L; \quad \beta_L = \mathbf{0}$$

(the choice of the reference class is arbitrary)

To each Y assign the corresponding indicator vector $\boldsymbol{\xi} \in \{0, 1\}^L$

MLE: $B \in \mathbb{R}^{d \times L}$ – matrix of regression coefficients ($B_{\cdot L} = \mathbf{0}$)

$$\ell(B) = \sum_{i=1}^n \left\{ \mathbf{x}_i^T B \boldsymbol{\xi}_i - \ln \sum_{l=1}^L \exp(\beta_l^T \mathbf{x}_i) \right\} \rightarrow \max_B$$

Sparse multinomial logistic regression

- for **multiclass** setup there are various ways to define sparsity
- **global row-wise sparsity**: part of features do not have any impact on classification at all, i.e. $B_j = \mathbf{0}$
- for a given model $M \subseteq \{1, \dots, d\}$

- ▶ $|M| = \#\{\text{non-zero rows of } B\} = r_B$



$$\hat{B}_M = \arg \max_{\tilde{B} \in \mathcal{B}_M} \sum_{i=1}^n \left\{ \mathbf{x}_i^T \tilde{B} \boldsymbol{\xi}_i - \ln \sum_{l=1}^L \exp(\tilde{\beta}_l^T \mathbf{x}_i) \right\},$$

where $\mathcal{B}_M = \{B \in \mathbb{R}^{d \times L} : B_{\cdot L} = \mathbf{0}, \text{ and } B_j = \mathbf{0} \text{ iff } j \notin M\}$

- $\hat{M} = \arg \min_M \{-\ell(\hat{B}_M) + \text{Pen}(|M|)\}$
- $\hat{\boldsymbol{\eta}}_{\hat{M}} = \arg \max_{1 \leq l \leq L} \hat{\beta}_{\hat{M}l}^T \mathbf{x}$

$$\mathcal{C}_L(d_0) = \{\eta(\mathbf{x}) = \arg \max_{1 \leq l \leq L} \beta_l^T \mathbf{x} : B \in \mathbb{R}^{d \times L}, B_{\cdot l} = \mathbf{0} \text{ and } r_B \leq d_0\}$$

Assumption (boundedness)

There exists $0 < \delta < 1/2$ such that $\delta \leq p_l(\mathbf{x}) \leq 1 - \delta$ or, equivalently, $|\beta_l^T \mathbf{x}| < C_0$ with $C_0 = \ln \frac{1-\delta}{\delta}$ for all $l = 1, \dots, L$ and $\mathbf{x} \in \mathcal{X}$.

Consider the complexity penalty

$$\text{Pen}(|M|) = C_1 \underbrace{|M|(L-1)}_{\# \text{ parameters, AIC}} + C_2 \underbrace{|M| \ln \left(\frac{de}{|M|} \right)}_{\log(\# \text{ models of size } |M|)}$$

Theorem (upper bound)

Assume d_0 -sparse multinomial logistic regression model. Under the boundedness assumption,

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\hat{\eta}_{\hat{M}}, \eta^*) \leq C(\delta) \sqrt{\frac{d_0(L-1) + d_0 \ln \left(\frac{de}{d_0} \right)}{n}}$$

Excess risk bounds

Theorem (lower bound)

Let $2 \leq d_0 \ln\left(\frac{de}{d_0}\right) \leq n$, $d_0(L-1) \leq n$ and $R(\eta^*) > 0$. Then,

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_L(d_0), f(\mathbf{x})} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C \sqrt{\frac{d_0(L-1) + d_0 \ln\left(\frac{de}{d_0}\right)}{n}}$$

The idea of the proof:

- 1 the error cannot be smaller than that for **binary** classification :

$$\text{Error} \geq C \sqrt{\frac{d_0 \ln\left(\frac{de}{d_0}\right)}{n}} \quad (\text{see above})$$

- 2 for a **given true (oracle)** model with $|M_0| = d_0$:

$$\text{Error} \geq C \sqrt{\frac{d_0(L-1)}{n}} \quad - \text{ via multiclass extension of VC (Daniely et al., '12, '15)}$$

Phase transition

- ① Small number of classes: $L \leq 2 + \ln \frac{d}{d_0}$
 - ▶ $Pen(|M|) \sim |M| \ln \frac{de}{|M|}$
 - ▶ the error is of the order $\sqrt{\frac{d_0 \ln \left(\frac{de}{d_0} \right)}{n}}$ (does not depend on L , binary case)
- ② Large number of classes: $2 + \ln \frac{d}{d_0} < L < \frac{n}{d_0}$
 - ▶ $Pen(|M|) \sim |M|(L - 1)$ (AIC)
 - ▶ the error is of the order $\sqrt{\frac{d_0(L-1)}{n}}$ (regardless of d)
- ③ $L > \frac{n}{d_0}$ – consistent classification is impossible

As before, the rates can be improved under the additional **low-noise condition** $P(p_{(1)}(\mathbf{X}) - p_{(2)}(\mathbf{X}) \leq h) \leq Ch^\alpha$, $\alpha \geq 0$, $0 < h < h^*$

Multinomial logistic **group** Lasso

B has a **global row-wise** sparsity. Let $|B|_j = \|B_j\|_2$

$$\widehat{B}_{gL} = \arg \min_{\widetilde{B}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\ln \left(\sum_{l=1}^L \exp(\widetilde{\beta}_l^T \mathbf{x}_i) \right) - \mathbf{x}_i^T \widetilde{B} \xi_i \right) + \lambda \sum_{j=1}^d |\widetilde{B}|_j \right\}$$

with $\lambda \sim \sqrt{\frac{L + \ln d}{n}}$

Let $n \gtrsim \ln d$. Under the boundedness and design assumptions,

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\widehat{\eta}_{gL}, \eta^*) \leq C(\delta) \sqrt{\frac{d_0(L-1) + d_0 \ln d}{n}}$$

(sub-optimal)

Multinomial logistic **group Slope**

$$\hat{B}_{gS} = \arg \min_{\tilde{B}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\ln \left(\sum_{l=1}^L \exp(\tilde{\beta}_l^T \mathbf{x}_i) \right) - \mathbf{x}_i^T \tilde{B} \xi_i \right) + \sum_{j=1}^d \lambda_j |\tilde{B}|_{(j)} \right\}$$

with $\lambda_j \sim \sqrt{\frac{L-1+\ln(de/j)}{n}}$

Let $n \gtrsim \ln d$. Under the boundedness and design assumptions,

$$\sup_{\eta^* \in \mathcal{C}_L(d_0)} \mathcal{E}(\hat{\eta}_{gS}, \eta^*) \leq C(\delta) \sqrt{\frac{d_0(L-1) + d_0 \ln \frac{de}{d_0}}{n}}$$

(optimal)

Extensions (in progress): Different types of sparsity

- **double row-wise sparsity**: nonzero rows are also sparse (local row-wise sparsity) – multinomial logistic **sparse group Slope**

$$\hat{B}_{sgS} = \arg \min_{\tilde{B}} \left\{ -\frac{1}{n} \ell(\tilde{B}) + \underbrace{\sum_{j=1}^d \lambda_j |\tilde{B}|_{(j)}}_{\text{group Slope}} + \underbrace{\sum_{j=1}^d \sum_{l=1}^L \kappa_l |\tilde{B}|_{j(l)}}_{\text{local (row-wise) Slope}} \right\}$$

with $\lambda_j \sim \sqrt{\frac{\ln(de/j)}{n}}$ and $\kappa_l \sim \sqrt{\frac{\ln(Le/l)}{n}}$

$$\mathcal{C}_L(d_0, \mathbf{m}) = \{\eta(\mathbf{x}) = \arg \max_{1 \leq l \leq L} \{\beta_l^T \mathbf{x} : |\mathcal{J}(B)| \leq d_0; \|B_j\|_0 \leq m_j, j \in \mathcal{J}(B)\}\}$$

Under the boundedness and design assumptions,

$$\sup_{\eta^* \in \mathcal{C}_L(d_0, \mathbf{m})} \mathcal{E}(\hat{\eta}_{sgS}, \eta^*) \leq C(\delta) \sqrt{\frac{d_0 \ln\left(\frac{de}{d_0}\right) + \sum_{j \in \mathcal{J}(B)} m_j \ln\left(\frac{Le}{m_j}\right)}{n}}$$

Extensions (in progress): Different types of sparsity

- **low rank sparsity:** B has zero singular values γ 's

$\text{rank}(B) = \|\gamma\|_0$ – not convex, use nuclear norm $\|B\|_* = \|\gamma\|_1$

$$\hat{B}_{nu} = \arg \min_{\tilde{B}} \left\{ -\frac{1}{n} \ell(\tilde{B}) + \lambda \|\tilde{B}\|_* \right\}$$

with $\lambda \sim \frac{\sqrt{L} + \sqrt{d}}{\sqrt{n}}$

$$\mathcal{C}_L(r_0) = \{\eta(\mathbf{x}) = \arg \max_{1 \leq I \leq L} \{\beta_I^T \mathbf{x} : \text{rank}(B) \leq r_0\}\}$$

Under the boundedness and design assumptions,

$$\sup_{\eta^* \in \mathcal{C}_L(r_0)} \mathcal{E}(\hat{\eta}_{nu}, \eta^*) \leq \sqrt{\frac{r_0(L+d)}{n}}$$

Thank You!