GEV-Canonical Regression for Accurate Binary Class Probability Estimation when One Class is Rare

Arpit Agarwal[1]

Harikrishna Narasimhan[1]

Shivaram Kalyanakrishnan[2]

Shivani Agarwal[1]

[1] **Indian Institute of Science, Bangalore**

[2] **Yahoo Labs, Bangalore**

Binary Problems where One Class is Rare

Fraud detection

Binary Problems where One Class is Rare

Fraud detection

Medical diagnosis

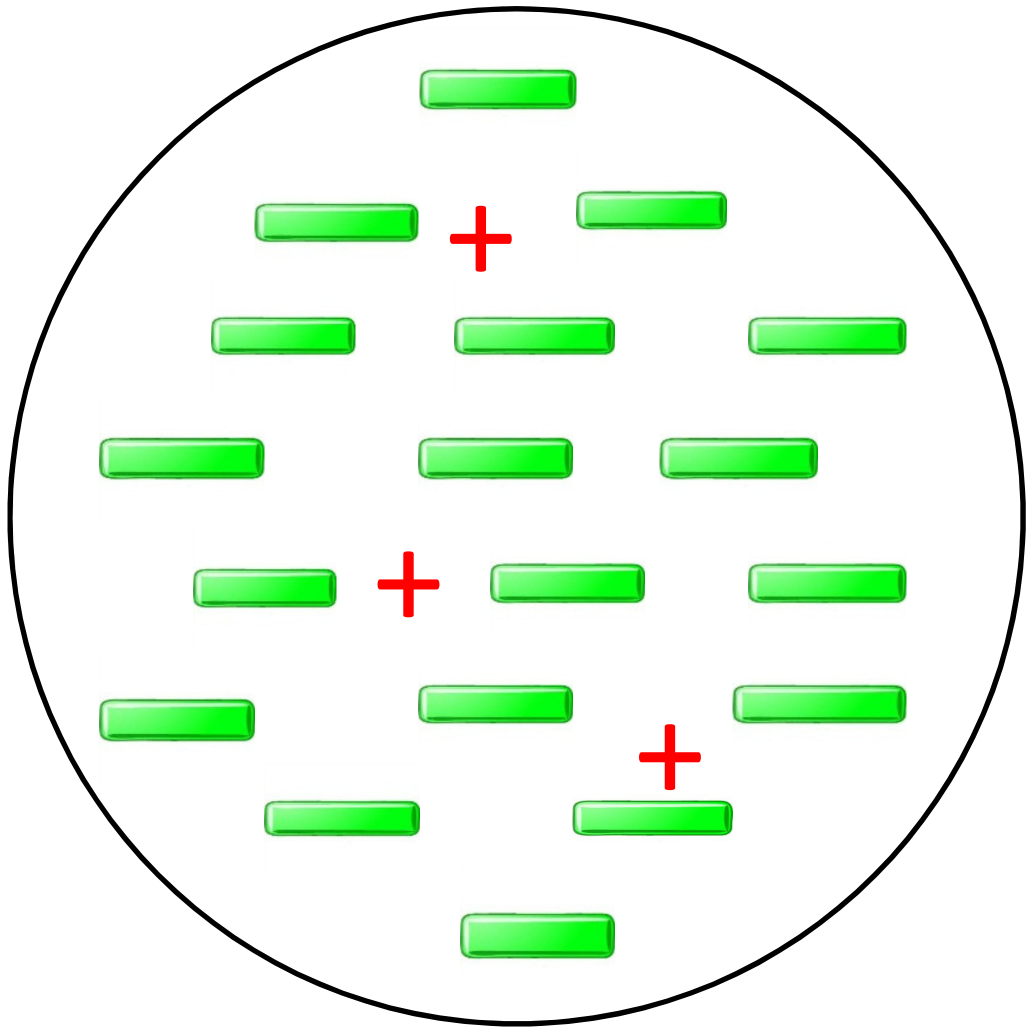
Binary Problems where One Class is Rare

Fraud detection

Medical diagnosis

Web-advertising

Binary Problems where One Class is Rare



Problem Setup

- Instance space $\mathcal{X}$, Label space $\mathcal{Y} = \{\pm 1\}$
- Probability distribution D on $\mathcal{X} \times \mathcal{Y}$
- $\eta(x) = \mathbf{P}(Y = 1 \,|\, X = x)$, $p = \mathbf{P}(Y = 1)$

Problem Setup

- Instance space $\mathcal{X}$, Label space $\mathcal{Y} = \{\pm 1\}$
- Probability distribution D on $\mathcal{X} \times \mathcal{Y}$
- $\eta(x) = \mathbf{P}(Y = 1 \,|\, X = x)$, $p = \mathbf{P}(Y = 1)$

We are interested in settings where $p \ll 0.5$

Problem Setup

- Instance space $\mathcal{X}$, Label space $\mathcal{Y} = \{\pm 1\}$
- Probability distribution D on $\mathcal{X} \times \mathcal{Y}$
- $\eta(x) = \mathbf{P}(Y = 1 \,|\, X = x)$, $p = \mathbf{P}(Y = 1)$

- Goal: Given a training sample

$$S = ((x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)) \sim D^n$$

learn a good class probability estimation (CPE) model $\widehat{\eta}_S : \mathcal{X} \to [0, 1]$

Previous Approaches

- Weighting errors on positive and negative examples differently (Provost, 2000; Japkowicz, 2000; Chawla et al., 2004; Van Hulse et al., 2007; He & Garcia, 2009)
- Undersampling majority class to balance positive and negative examples (King & Zeng, 2001)
- Asymmetric \`link' function based on generalized extreme value (GEV) distribution (Wang & Dey, 2010; Calabrese & Osmetti, 2011)

Our Work

- We use tools from the theory of proper composite losses to design a loss based on the GEV link termed GEV-canonical
- GEV-canonical loss is both flexible and convex
- We also propose the GEV-canonical regression algorithm for its minimization

Outline

- Proper Composite Loss Functions
- GEV-Canonical Loss Function &
  GEV-Canonical Regression Algorithm
- Experiments

Loss Functions for CPE

- A CPE loss function $c : \{\pm 1\} \times [0,1] \to \overline{\mathbb{R}}_+$ assigns a penalty $c(y, \widehat{\eta})$ for predicting $\widehat{\eta}$ when the true label is y

Loss Functions for CPE

- A CPE loss function $c : \{\pm 1\} \times [0,1] \rightarrow \overline{\mathbb{R}}_+$ assigns a penalty $c(y, \widehat{\eta})$ for predicting $\widehat{\eta}$ when the true label is y
- Can be defined by its partial losses $c_1 : [0,1] \rightarrow \overline{\mathbb{R}}_+$ and $c_{-1} : [0,1] \rightarrow \overline{\mathbb{R}}_+$, given by

$$c_y(\widehat{\eta}) = c(y, \widehat{\eta})$$

Proper Loss Functions

A CPE loss function $c : \{\pm 1\} \times [0,1] \rightarrow \overline{\mathbb{R}}_+$ is proper if

$$\eta \in \arg\min_{\widehat{\eta} \in [0,1]} \eta \, c_1(\widehat{\eta}) + (1-\eta) \, c_{-1}(\widehat{\eta}) \qquad \forall \eta \in [0,1]$$

and strictly proper if the minimizer is unique

Example: Logarithmic Loss

$$\begin{aligned} c_1^{\log}(\widehat{\eta}) &= -\ln(\widehat{\eta})\,; \\ c_{-1}^{\log}(\widehat{\eta}) &= -\ln(1-\widehat{\eta})\,. \end{aligned}$$

Example: Logarithmic Loss

$$
\begin{aligned}
c_1^{\log}(\widehat{\eta}) &= -\ln(\widehat{\eta})\,; \\
c_{-1}^{\log}(\widehat{\eta}) &= -\ln(1-\widehat{\eta})\,.
\end{aligned}
$$

$$
\eta = \arg\min_{\widehat{\eta}\in[0,1]} [-\eta\ \ln(\widehat{\eta}) - (1-\eta)\ \ln(1-\widehat{\eta})]
$$

Log loss is strictly proper

Link Functions

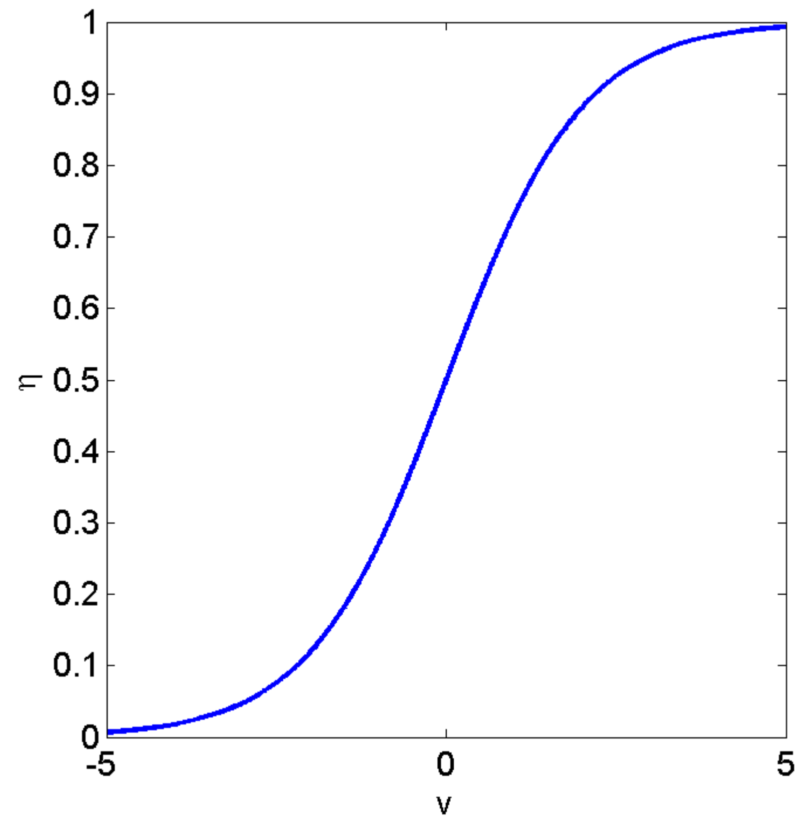
Let $\mathcal{V} \subseteq \mathbb{R}$, A link function

$$\psi : [0, 1] \rightarrow \mathcal{V}$$

is any strictly increasing (and therefore invertible) function that maps probabilities in $[0, 1]$ to real-valued scores in $\mathcal{V}$

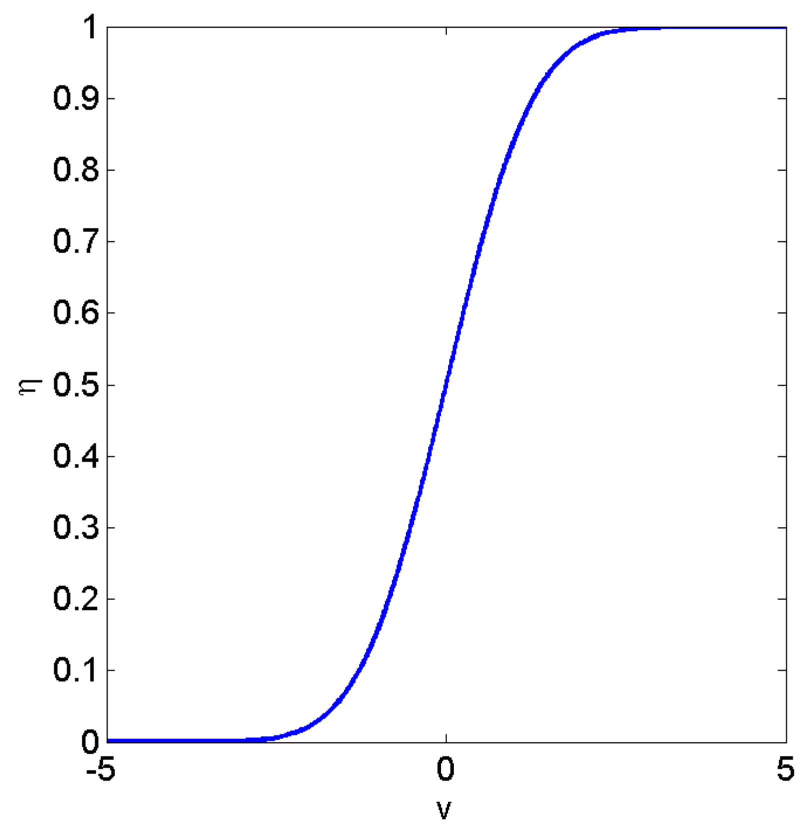
Example: Logit Link

$$\psi_{\text{logit}}(\widehat{\eta}) = \ln\left(\frac{\widehat{\eta}}{1-\widehat{\eta}}\right)$$



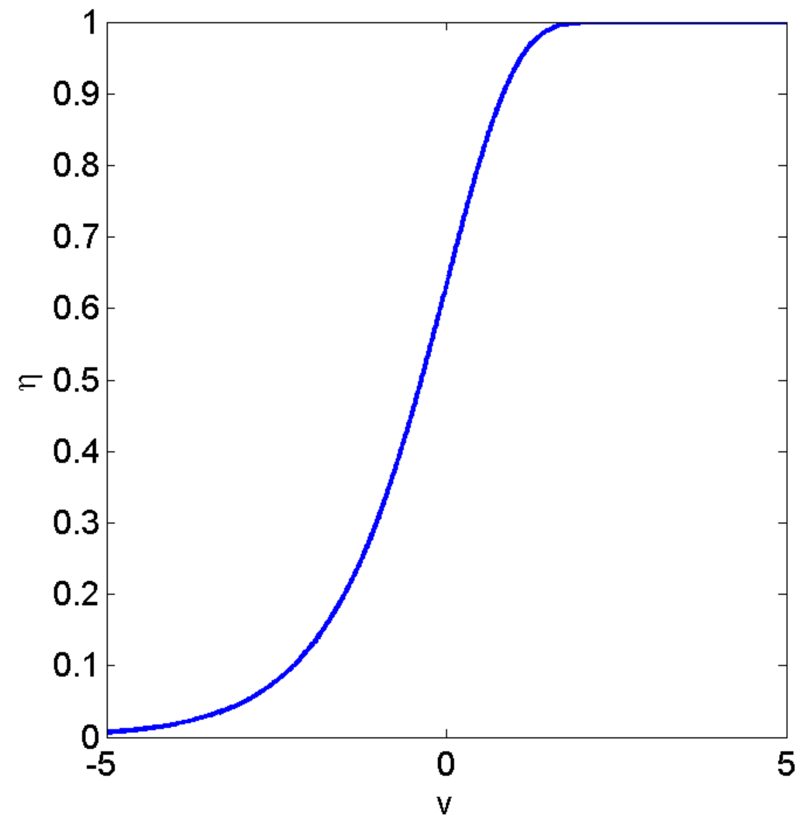
Example: Probit Link

$$\psi_{\text{probit}}(\widehat{\eta}) = \Phi^{-1}(\widehat{\eta})$$



Example: Complementary Log-Log Link

$$\psi_{\text{cloglog}}(\widehat{\eta}) = \ln(-\ln(1 - \widehat{\eta}))$$



Proper Composite Loss Functions

[Buja et al, 2005; Reid & Williamson, 2009, 2010]

A loss function $\ell : \{\pm 1\} \times \mathcal{V} \rightarrow \overline{\mathbb{R}}_+$ is said to be proper composite if $\exists$ a proper CPE loss $c : \{\pm 1\} \times [0,1] \rightarrow \overline{\mathbb{R}}_+$ and a link $\psi : [0,1] \rightarrow \mathcal{V}$ s.t.

$$\ell(y, v) = c(y, \psi^{-1}(v))$$

Canonical Proper Loss & Link Pairs

[Buja et al, 2005; Reid & Williamson, 2009, 2010]

- For every link function ψ there is a unique canonical proper loss function given by:

$$
\begin{aligned}
c_1(\widehat{\eta}) &= \int_{\widehat{\eta}}^{1} (1-q)\,\psi'(q)\,dq\,; \\
c_{-1}(\widehat{\eta}) &= \int_{0}^{\widehat{\eta}} q\,\psi'(q)\,dq\,,
\end{aligned}
$$

Canonical Proper Loss & Link Pairs

[Buja et al, 2005; Reid & Williamson, 2009, 2010]

- For every link function ψ there is a unique canonical proper loss function given by:

$$
\begin{aligned}
c_1(\widehat{\eta}) &= \int_{\widehat{\eta}}^{1} (1-q)\,\psi'(q)\,dq\,; \\
c_{-1}(\widehat{\eta}) &= \int_{0}^{\widehat{\eta}} q\,\psi'(q)\,dq\,,
\end{aligned}
$$

- The resulting proper composite loss has some nice properties, including convexity.

Example: Logistic Loss

Log Loss + Logit Link = Logistic Loss

$$\ell^{\text{logistic}}(y, v) \;=\; -\ln\left(\frac{1}{1+\exp(-yv)}\right)$$

Example: Logistic Loss

Log Loss + Logit Link = Logistic Loss

$$\ell^{\text{logistic}}(y, v) = -\ln\left(\frac{1}{1+\exp(-yv)}\right)$$

Canonical pair

Outline

- Proper Composite Loss Functions
- GEV-Canonical Loss Function &
  GEV-Canonical Regression Algorithm
- Experiments

Generalized Extreme Value (GEV) Probability Distribution

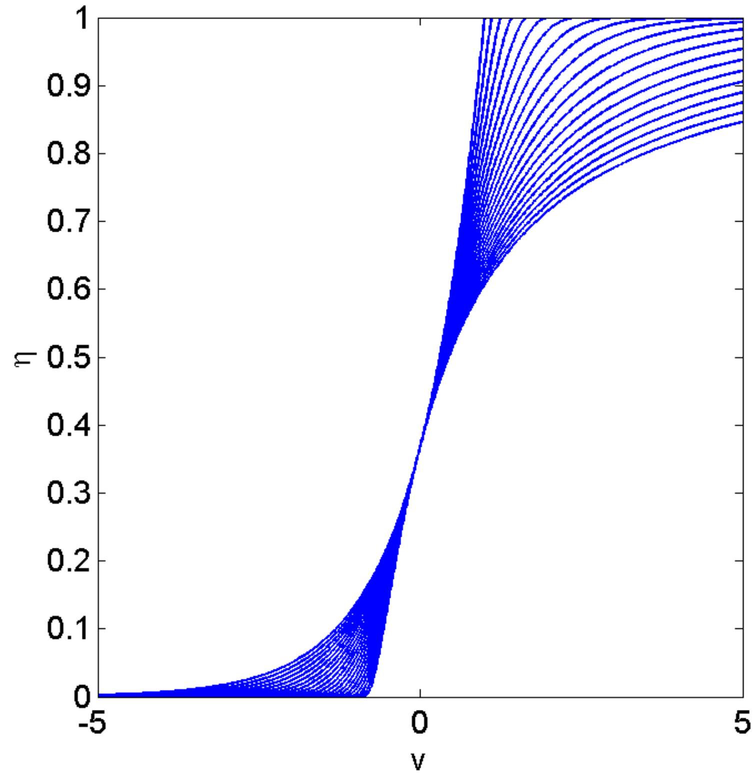
- CDF of GEV distribution with location parameter $\mu = 0$, scale parameter $\sigma = 1$, and shape parameter $\xi \in \mathbb{R}$:

$$F_\xi(v) = \exp(-(1 + \xi v)_+^{-1/\xi}).$$

- Used for modeling rare events in statistics

GEV Link Family (Parameterized by $\xi \in \mathbb{R}$)

$$\psi_{\mathrm{GEV}(\xi)}(\widehat{\eta}) = \frac{1}{\xi}\left(\frac{1}{\left(-\ln(\widehat{\eta})\right)^{\xi}} - 1\right)$$



GEV-Log Loss Effectively Used in
(Wang & Dey, 2010; Calabrese & Osmetti, 2011)

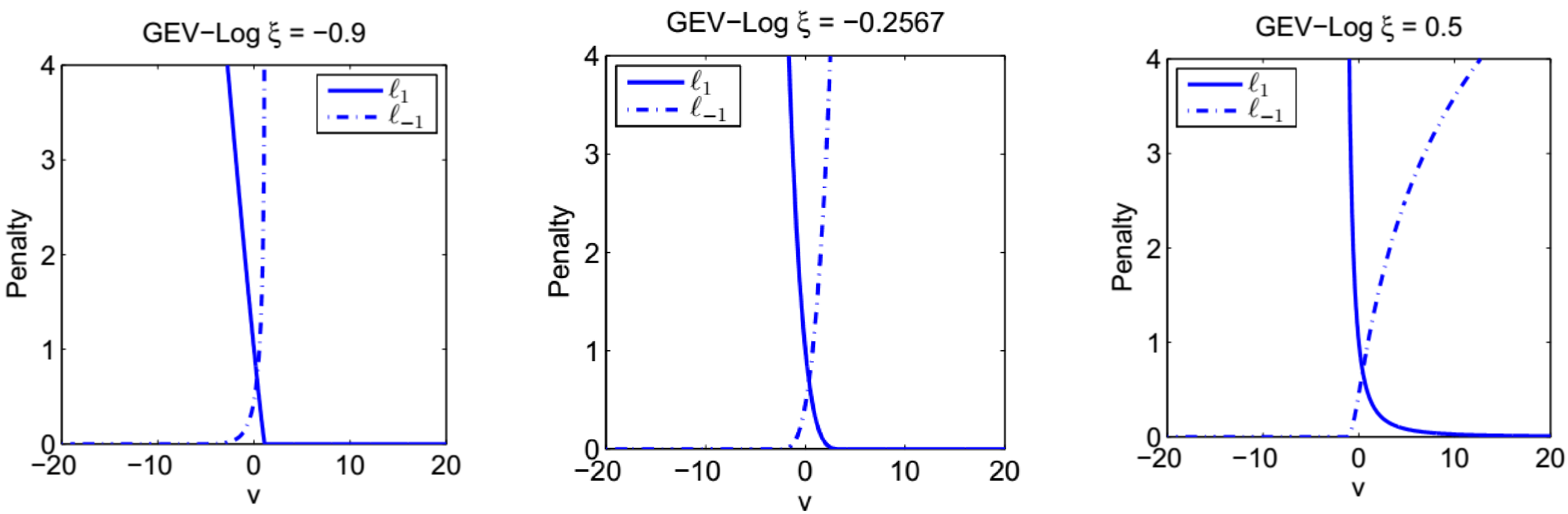
Log Loss + GEV Link = GEV-Log Loss

$$\ell^{\text{GEV-}\log(\xi)}(y,v) = -\mathbf{1}[y=1]\ \ln(\psi_{GEV}^{-1}(v;\xi)) - \mathbf{1}[y=-1]\ \ln(1-\psi_{GEV}^{-1}(v;\xi))$$

GEV-Log Loss Effectively Used in
(Wang & Dey, 2010; Calabrese & Osmetti, 2011)

Log Loss + GEV Link = GEV-Log Loss

$$\ell^{\text{GEV- log}(\xi)}(y, v) = -\mathbf{1}[y = 1] \ln(\psi^{-1}_{GEV}(v;\xi)) - \mathbf{1}[y = -1] \ln(1 - \psi^{-1}_{GEV}(v;\xi))$$

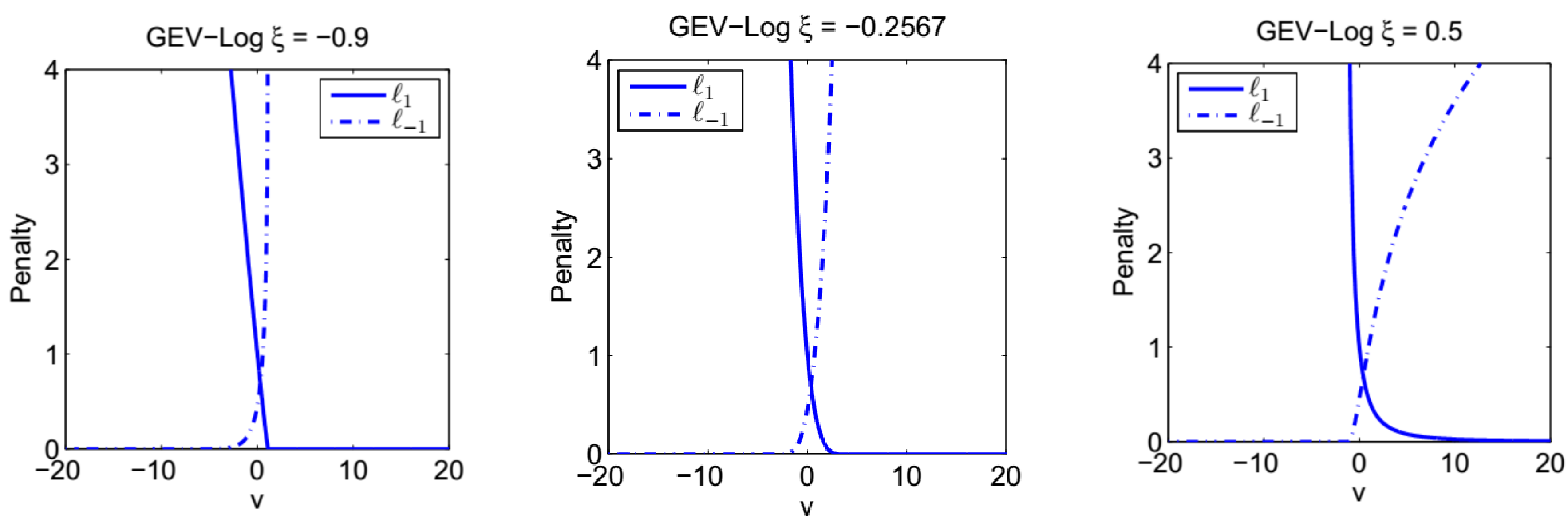


GEV-Log Loss Effectively Used in
(Wang & Dey, 2010; Calabrese & Osmetti, 2011)

Log Loss + GEV Link = GEV-Log Loss

$$\ell^{\text{GEV-}\log(\xi)}(y,v) = -\mathbf{1}[y=1]\,\ln(\psi_{GEV}^{-1}(v;\xi)) - \mathbf{1}[y=-1]\,\ln(1-\psi_{GEV}^{-1}(v;\xi))$$

NOT a canonical pair; results in non-convex loss



Canonical Proper Loss for GEV Link

$$
\begin{aligned}
c_1^{\text{GEV-can}(\xi)}(\widehat{\eta}) &= \int_{\widehat{\eta}}^{1} \frac{1-q}{q(-\ln q)^{1+\xi}}\, dq \\
c_{-1}^{\text{GEV-can}(\xi)}(\widehat{\eta}) &= \int_{0}^{\widehat{\eta}} \frac{1}{(-\ln q)^{1+\xi}}\, dq
\end{aligned}
$$

GEV-Canonical Loss

(Canonical Loss) + GEV Link = GEV-Canonical Loss

$$\ell^{\text{GEV-can}(\xi)}(y,v) = -\mathbf{1}[y=1]\, c_1^{\text{GEV-can}(\xi)}(\psi_{\text{GEV}}^{-1}(v;\xi)) - \mathbf{1}[y=-1]\, c_{-1}^{\text{GEV-can}(\xi)}(1-\psi_{\text{GEV}}^{-1}(v;\xi))$$

GEV-Canonical Loss

(Canonical Loss) + GEV Link = GEV-Canonical Loss

$$\ell^{\text{GEV-can}(\xi)}(y,v) = -\mathbf{1}[y=1]\, c_1^{\text{GEV-can}(\xi)}(\psi_{\text{GEV}}^{-1}(v;\xi)) - \mathbf{1}[y=-1]\, c_{-1}^{\text{GEV-can}(\xi)}(1-\psi_{\text{GEV}}^{-1}(v;\xi))$$

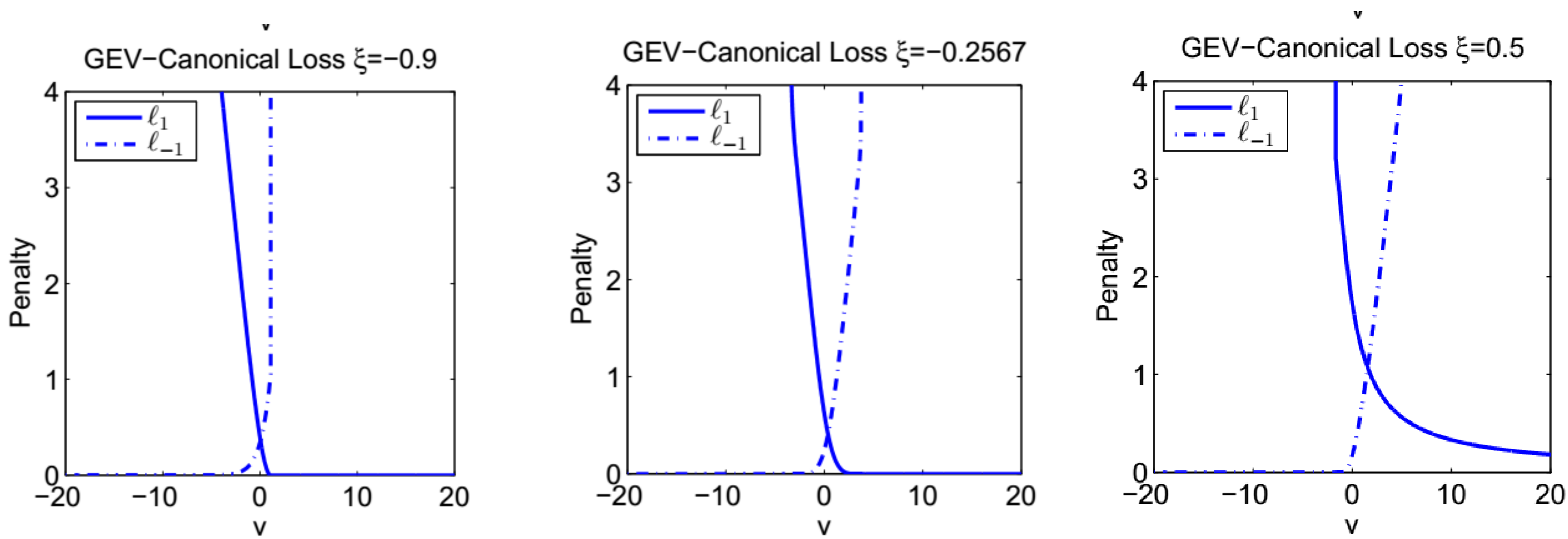
Canonical pair by construction; results in convex loss!

GEV-Canonical Loss

(Canonical Loss) + GEV Link = GEV-Canonical Loss

$$\ell^{\text{GEV-can}(\xi)}(y,v) = -\mathbf{1}[y=1]\, c_1^{\text{GEV-can}(\xi)}(\psi_{\text{GEV}}^{-1}(v;\xi)) - \mathbf{1}[y=-1]\, c_{-1}^{\text{GEV-can}(\xi)}(1-\psi_{\text{GEV}}^{-1}(v;\xi))$$

Canonical pair by construction; results in convex loss!



GEV-Canonical Loss

- Can be tailored for the problem of CPE for varying degrees of rarity
- Not available in closed form. But, the gradient and Hessian are available in closed form
- Can be efficiently minimized using IRLS type algorithm. We term this GEV-canonical regression

GEV-Canonical Regression

Algorithm 1 GEV-Canonical Regression (using IRLS)

Input: Data $S = ((\mathbf{x}_i, y_i))_{i=1}^n \in (\mathbb{R}^k \times \{\pm 1\})^n$

Initialize: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times k}$

$$\eta_i^{(1)} = \begin{cases} 0.75 & \text{if } y_i = 1 \\ 0.25 & \text{if } y_i = -1 \end{cases} \quad \forall i \in [n]$$

$$v_i^{(1)} = \psi_{\text{GEV}(\xi)}(\eta_i^{(1)}) \quad \forall i \in [n]$$

$$t = 1$$

repeat

  for $i = 1$ **to** n **do**

    $w_i^{(t)} = \eta_i^{(t)}(-\ln(\eta_i^{(t)}))^{\xi+1}$

    choose a suitable step size $\gamma^{(t)}$

    $z_i^{(t)} = v_i^{(t)} + \gamma^{(t)} \cdot \big(\mathbf{1}(y_i = 1) - \eta_i^{(t)}\big) \cdot \psi'_{\text{GEV}(\xi)}(\eta_i^{(t)})$

  end for

  $\mathbf{W}^{(\mathbf{t})} = \text{diag}(w_1^{(t)}, \cdots, w_n^{(t)})$

  $\boldsymbol{\beta}^{(t)} = (\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{z}^{(t)}$

    // compute $\boldsymbol{\beta}^{(t)}$ via weighted least squares (WLS)

  for $i = 1$ **to** n **do**

    $v_i^{(t+1)} = (\boldsymbol{\beta}^{(t)})^\top \mathbf{x}_i$

    $\eta_i^{(t+1)} = \psi_{\text{GEV}(\xi)}^{-1}(\text{clip}_\xi(v_i^{(t+1)}))$

  end for

  $t \leftarrow t + 1$

until convergence

Output: Coefficient vector $\boldsymbol{\beta}^{(t-1)} \in \mathbb{R}^k$

Outline

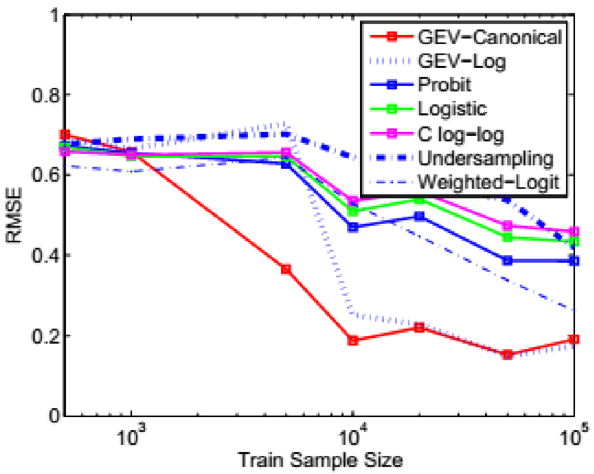
- Proper Composite Loss Functions
- GEV-Canonical Loss Function &
  GEV-Canonical Regression Algorithm
- Experiments

Experiments

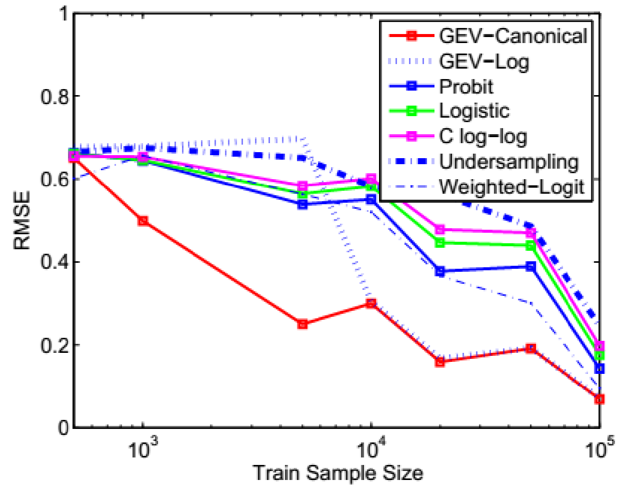
- We have conducted experiments with both synthetic and real data
- Parameter ξ selected using a validation set.
- Results averaged over 10 experiments.

Experiments with Synthetic Data

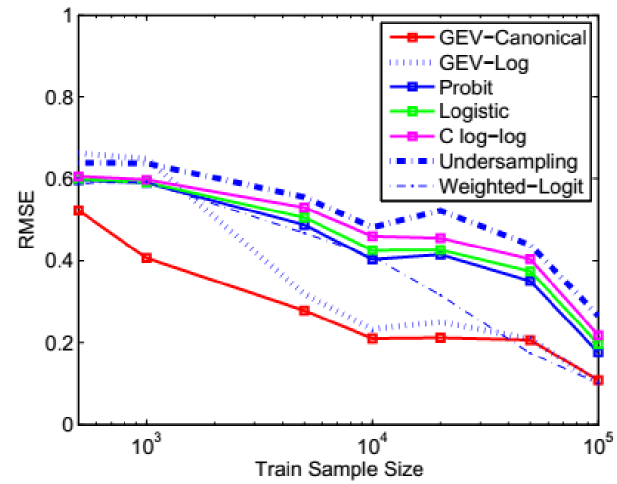
- Evaluation Metric: Root Mean Square Error (RMSE)



Dataset 1 : $p = 0.0158$



Dataset 2 : $p = 0.0312$



Dataset 3 : $p = 0.095$

Experiments with Real Data

- Experimented with 12 UCI data sets
- Evaluation Metric: Brier Score (Brier, 1950)

Dataset	Logistic regression	Probit regression	Cloglog regression	Undersampling + King-Zeng correction	Weighted Logistic + correction	GEV-log regression	GEV-canonical regression
Nursery	**0.0084**	**0.0084**	0.0088 **	0.0124 **	0.0090 **	0.0172 **	**0.0084**
Letter-a	0.0079 **	0.0084 **	**0.0074**	0.0111 **	0.0112 **	0.0313 **	0.0080 **
Car	0.0266 **	0.0262	0.0267 **	0.0320 **	0.0271 *	0.0296 **	**0.0259**
Glass	0.0670	0.0671	0.0744 **	0.0623	0.0637	**0.0614**	0.0649
Ecoli	0.0646	0.0644	0.0689 **	0.0756 **	**0.0635**	0.0641	0.0641
Letter-vowel	0.1392 **	0.1392 **	0.1400 **	0.1416 **	0.1414 **	0.1405 **	**0.1367**
Cmc	0.1617	0.1617	0.1621	0.1642 **	**0.1615**	0.1626	0.1622
Vehicle	0.1399	0.1395	0.1422	0.1501 **	0.1408	0.1497 **	**0.1394**
Haberman	0.1828 *	0.1812	0.1907 **	0.1823 *	0.1814 **	**0.1761**	0.1769
Yeast	0.1634 **	0.1635 **	0.1666 **	0.1646 **	0.1635 **	0.1621	**0.1616**
German	0.1721	0.1731	0.1737	0.1754 **	**0.1714**	0.1787 **	0.1727
Pima	0.1617 **	0.1623 **	0.1652 **	0.1662 *	0.1626 **	0.1616	**0.1603**

Summary

	FLEXIBLE?	CONVEX?
LOGISTIC	$\times$	$\surd$
PROBIT	$\times$	$\surd$
CLOGLOG	$\times$	$\surd$
GEV-LOG(ξ)	$\surd$	$\times$
GEV-CANONICAL(ξ)	$\surd$	$\surd$

Conclusion and Future Work

- Proposed GEV-canonical regression algorithm using convex GEV-canonical loss for the problem of CPE when one class is rare
- Future directions:
  - extensions to large scale data
  - statistical guarantees