Fenchel-Young Losses with Skewed Entropies for Class-posterior Probability Estimation

Han Bao[1,2], Masashi Sugiyama[2,1]
[1]The University of Tokyo [2]RIKEN, Japan

Summary

- Problem: model binary outcomes, specifically estimate $\mathbb{P}(\mathrm{Y}=1\mid \boldsymbol{x})$
- Common approach: logistic regression
  - ☺ identifiability of model parameter
  - ☹ link misspecification due to symmetry of link
- Existing remedy: replace logit link with more flexible link family
  - ☺ resolve link misspecification by model selection
  - ☹ maximum log-likelihood is no longer convex
- **Proposal: Fenchel-Young loss + flexible link = convex loss**

Implementation available: http://bit.ly/gh-GEV-FY

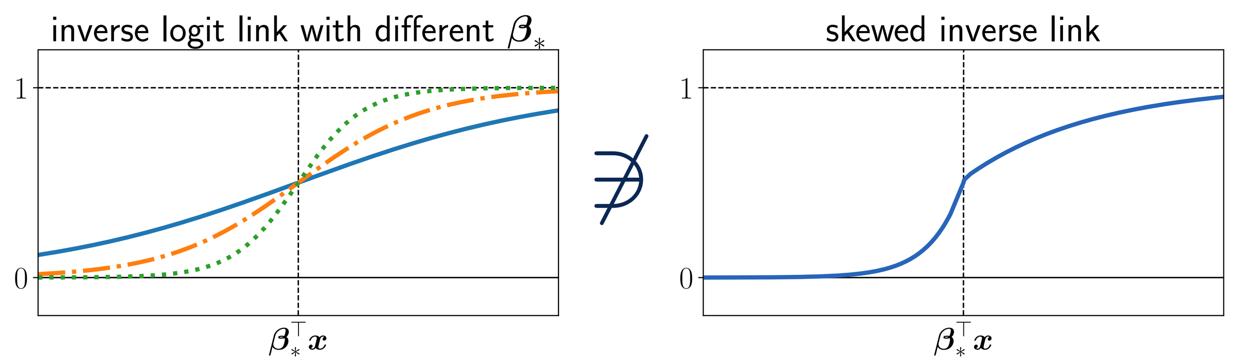
Introduction

Logistic regression

$$\mathrm{Y}=1\mid \boldsymbol{x}\sim \mathrm{Bernoulli}(\eta) \quad \text{where} \quad \eta = \underbrace{\psi^{-1}(\boldsymbol{\beta}_*^\top \boldsymbol{x})}_{\text{inverse link}} = \frac{1}{1+e^{-\boldsymbol{\beta}_*^\top \boldsymbol{x}}}$$

Widely used for modeling binary outcomes (e.g. epidemiology), where $\psi^{-1}(\boldsymbol{\beta}^\top \boldsymbol{x})$ models $\mathbb{P}(\mathrm{Y}=1\mid \mathrm{X}=\boldsymbol{x})$, but unable to accommodate skewed link functions (e.g. class imbalance)

$\Longrightarrow$ Needs more flexible link!

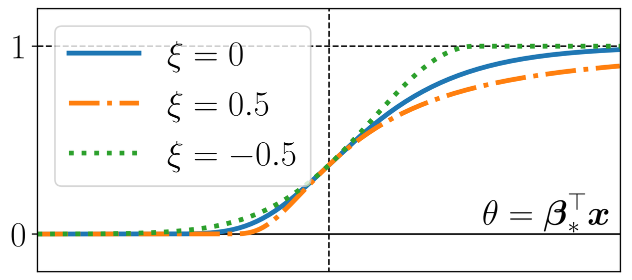


GEV (generalized extreme value) link family [1]

$$\psi^{-1}(\theta)=\exp\left(\left(1+\xi\theta\right)_+^{-1/\xi}\right)$$

(ξ: shape parameter)

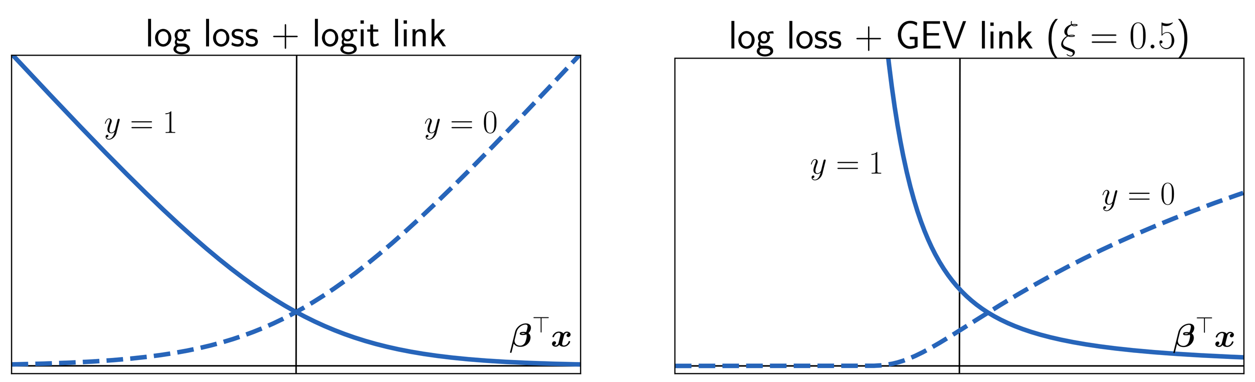
one parameter ξ controls skewness of link, which can be chosen via model selection



Fitting: MLE with logit link is convex, while GEV link results in a non-convex problem

$$\min_{\boldsymbol{\beta}} \sum_i \underbrace{-\log}_{\text{log loss}} \psi^{-1}(\boldsymbol{\beta}^\top \boldsymbol{x}_i)^{y_i}(1-\psi^{-1}(\boldsymbol{\beta}^\top \boldsymbol{x}_i))^{1-y_i}$$

Q. Any convex loss for GEV link?



Background: Fenchel-Young Loss [2]

A framework to generate a loss from an entropic regularizer of *prediction function*, which maps logit $\theta\in\mathbb{R}$ to probabilistic prediction $\widehat{y}_\Omega(\theta)\in[0,1]$:

$$\widehat{y}_\Omega(\theta) = \underbrace{\operatorname*{arg\,max}_{\eta\in[0,1]} \theta\eta}_{1 \text{ if } \eta>1/2 \text{ otherwise } 0} - \underbrace{\Omega(\eta)}_{\text{entropy}}$$

Fenchel-Young loss

Let $\Omega:[0,1]\to\mathbb{R}$ be a regularizer, $y\in\{0,1\}$ be a label, and $\theta\in\mathbb{R}$ be a logit score. Then, *Fenchel-Young loss* $\ell_\Omega(\theta;y)$ generated by Ω is

$$\ell_\Omega(\theta;y)\stackrel{\text{def}}{=}\Omega^*(\theta)+\Omega(y)-\theta y; \qquad \text{where} \quad \Omega^*(\theta)=\sup_{\eta\in[0,1]}\theta\eta-\Omega(\eta).$$

Property

- Convexity in θ
- Zero-loss: $\ell_\Omega(\theta;y)=0 \Leftrightarrow y=\widehat{y}_\Omega(\theta)$

Example

Ω	ℓ_Ω	$\widehat{y}_\Omega$
Shannon	logistic	softmax
2-Tsallis	modified Huber	sparsemax

Q. How to derive a regularizer Ω that we desire?

Our Idea: Generate Loss from Link

A. Integrate link function ψ to derive Ω

by identifying inverse link ψ^{-1} and prediction function $\widehat{y}_\Omega$

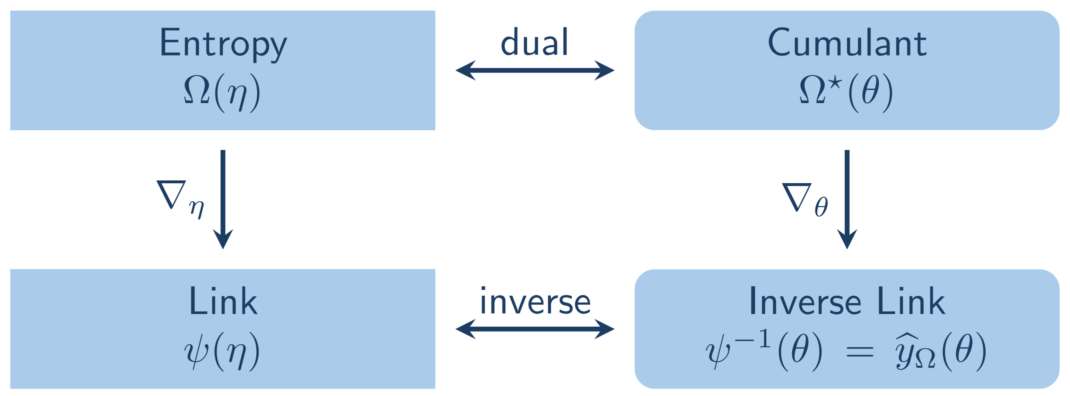


Figure 1: The relationship obtained thanks to convex analysis and Danskin's theorem

In case of GEV link,

$$\Omega(\eta)=\int_0^\eta \psi(\eta)\mathrm{d}\eta=\frac{1}{\xi}\left(\Gamma(1-\xi,-\log\eta)-\eta\right),$$

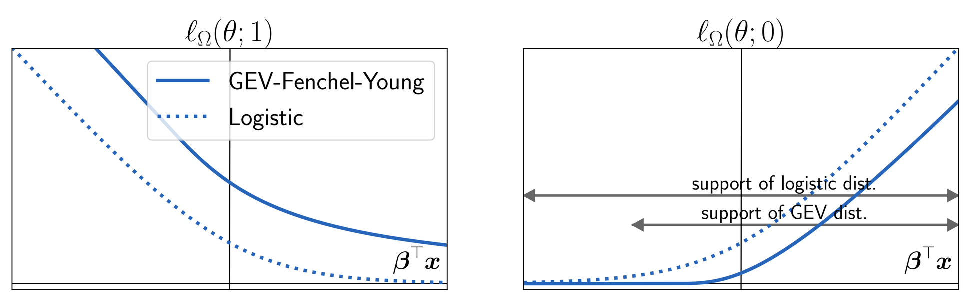
$$\Omega^*(\theta)=\int_{-\infty}^\theta \psi^{-1}(\theta)\mathrm{d}\theta=\begin{cases}\Gamma(-\xi,(1+\xi\theta)^{-1/\xi}) & \text{if } \theta\le -1/\xi\\ \theta+\Gamma(-\xi,0)+\xi^{-1} & \text{if } \theta>-1/\xi.\end{cases}$$

($\xi<1$, Γ is incomplete Gamma function)

Good property: partial separation margin

GEV Fenchel-Young loss ($\xi>0$) attains $\ell_\Omega(\theta;y=0)=0$ with some finite logit

$\Longrightarrow$ penalize logit of $y=1$ (rare class) heavier hence beneficial for class imbalance



ℓ_Ω has this property if support of ψ^{-1} is bounded at left end

Comparison with Proper Loss

Canonical proper composite loss is another framework to ensure loss convexity

Proper composite loss [3]

- loss $\ell(\widehat{\eta};\eta)$ is *proper* if $L(\eta;\eta)=\underline{L}(\eta)$ $\left(\widehat{\eta},\eta\in[0,1]\right)$
  ($L(\widehat{\eta};\eta)=\mathbb{E}_{\mathrm{Y}\sim\eta}[\ell(\widehat{\eta};\eta)]$: conditional risk, $\underline{L}(\eta)=\inf_{\widehat{\eta}}L(\widehat{\eta};\eta)$: Bayes risk)
- $\ell(\psi^{-1}(\theta);\eta)$ is *proper composite* for inverse link ψ^{-1} and proper loss ℓ
- (ℓ,ψ) is a *canonical* pair if $\psi=-\nabla\underline{L}$
- $\ell(\psi^{-1}(\theta;\eta))$ is convex in $\theta\in\mathrm{Im}(\psi)$ if (ℓ,ψ) is canonical

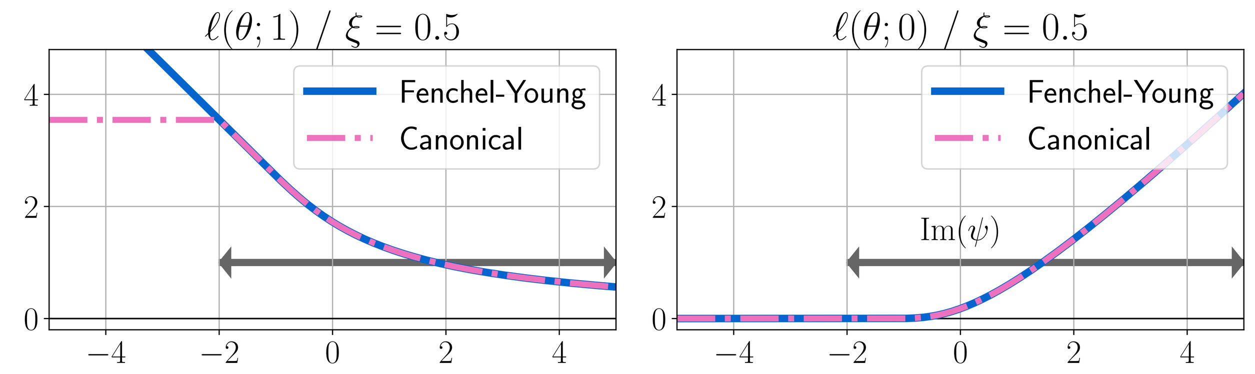


Figure 2: Comparison of canonical loss and Fenchel-Young loss generated from GEV link

- Fenchel-Young loss matches canonical proper loss for $\theta\in\mathrm{Im}(\psi)$
- Canonical proper loss is no longer convex in $\theta\in\mathbb{R}$

$\Longrightarrow$ Fenchel-Young loss systematically extrapolate canonical proper loss!
(which is convenient for flexible links such as $\mathrm{Im}(\psi)\ne\mathbb{R}$)

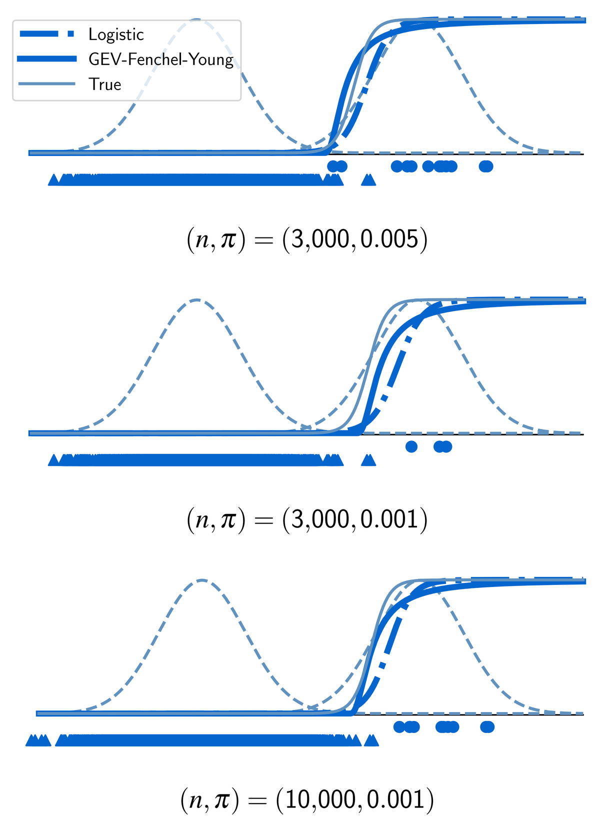
Simulation

Setup: compare logit and GEV link with different $\pi=\mathbb{P}(\mathrm{Y}=1)$ and sample size n

- Data: $\mathrm{X}\mid\mathrm{Y}=y\sim\mathcal{N}(2y-1,0.4)$
- Optimization: 100 epochs with Adam (lr $=1$)
- ξ is fixed to 0.5

Result: GEV-Fenchel-Young loss is tolerant to heavy imbalance ($\pi=0.001$) with large enough samples ($n=10{,}000$), while logistic loss is still biased (see bottom figure)

Remark: larger experiments and F-measure optimization are performed as well in the paper



References

[1] Wang, X. and Dey, D. K. (2010). Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *The Annals of Applied Statistics.*

[2] Blondel, M., Martins, A. F., and Niculae, V. (2020). Learning with Fenchel-Young losses. *Journal of Machine Learning Research.*

[3] Reid, M. D. and Williamson, R. C. (2010). Composite binary losses. *Journal of Machine Learning Research.*