Generalization Bounds and Consistency for Latent-Structural Probit and Ramp Loss

David McAllester and Joseph Keshet

Toyota Technological Institute at Chicago
Why Surrogate Loss Functions

We consider an arbitrary input space $\mathcal{X}$ and a finite label space $\mathcal{Y}$, a source probability distribution over pairs $(x, y)$, and a task loss $L$ with $L(y, \hat{y}) \in [0, 1]$. We will use a linear classifier with parameter vector $w$

$$\hat{y}_w(x) = \arg\max_y w^\top \phi(x, y)$$

We would like

$$w^* = \arg\min_w \mathbb{E}_{x,y} [L(y, \hat{y}_w(x))]$$

We get

$$\hat{w} = \arg\min_w \left( \sum_{i=1}^{n} L_s(w, x_i, y_i) \right) + \frac{\lambda}{2} ||w||^2$$

$L_s$ must be scale-sensitive and hence different from $L$. 

Structural Surrogate Loss Functions

\[ L_{\log}(w, x, y) = \ln \frac{1}{P_w(y|x)} \]

\[ L_{\text{hinge}}(w, x, y) = \left( \max_{\hat{y}} w^\top \phi(x, \hat{y}) + L(y, \hat{y}) \right) - w^\top \Phi(x, y) \]

\[ L_{\text{ramp}}(w, x, y) = \left( \max_{\hat{y}} w^\top \phi(x, \hat{y}) + L(y, \hat{y}) \right) - \left( \max_{\hat{y}} w^\top \Phi(x, \hat{y}) \right) \]

\[ L_{\text{probit}}(w, x, y) = \mathbb{E}_\epsilon [L(y, \hat{y}_w + \epsilon(x))] \]
Surrogate Loss in the Binary Case

\[ y \in \{-1, 1\}, \quad L(y, \hat{y}) = 1_{y \neq \hat{y}} \]

\[ \phi(x, y) = y\phi(x)/2, \quad m = yw^\top \phi(x), \]

\[ L_{\log}(w, x, y) = \ln(1 + e^{-m}) \]

\[ L_{\text{hinge}}(w, x, y) = \max(0, 1 - m) \]

\[ L_{\text{ramp}}(w, x, y) = \min(1, \max(0, 1 - m)) \]

\[ L_{\text{probit}}(w, x, y) = P_{\epsilon \sim \mathcal{N}(0, 1)}[\epsilon \geq m] \quad \text{for} \quad \|\Phi(x)\| = 1 \]
Basic Properties

Both $L_{\log}(w, x, y)$ and $L_{\text{hinge}}(w, x, y)$ increase without bound when scaling up $w$ on a wrongly classified training point.

$L_{\text{ramp}}(w, x, y), L_{\text{probit}}(w, x, y) \in [0, 1]$

$L_{\text{hinge}}(w, x, y) \geq L_{\text{ramp}}(w, x, y) \geq L(w, x, y)$

The last relation motivates $L_{\text{ramp}}$ in [Do, Le, Teo, Chapelle, and Smola, 2008].
Empirical Studies

Subgradient descent on unregularized ramp loss (and related methods) have been shown to give improvements over hinge loss in machine translation and speech applications.

- D. Chiang, K. Knight, and W. Wang. NAACL, 2009
- Keshet, Cheng, Stoehr, and McAllester, Interspeech 2011

Probit loss has been show to give an improvement over hinge loss for phonetic transcription.

- Keshet, McAllester, and Hazan, ICASSP, 2011.
### Some Notation

\[ L(w) = \mathbb{E}_{x,y} [L(w, x, y)] \]

\[ L^* = \inf_w L(w) \]

\[ \hat{L}^n(w) = \frac{1}{n} \sum_{i=1}^{n} L(w, x_i, y_i) \]
Consistency of Probit Loss

We consider the following learning rule where $\lambda_n$ is some given function of $n$.

\[ \hat{w}_n = \arg\min_w \hat{L}^n_{\text{probit}}(w) + \frac{\lambda_n}{2n} ||w||^2 \]

If

- $\lambda_n$ increases without bound
- $(\lambda_n \ln n)/n$ converges to zero

then

\[ \lim_{n \to \infty} L_{\text{probit}}(\hat{w}_n) = L^* \]
**PAC-Bayesian Bounds**

[Catoni 07], [Germain, Lacasse, Laviolette, Marchand 09]

For $L(\hat{y}, y) \in [0, 1]$, and for any fixed prior distribution $P$ and fixed $\lambda > 1/2$ we have that with probability at least $1 - \delta$ over the draw of the training data the following holds simultaneously for all $Q$.

$$L(Q) \leq \frac{1}{1 - \frac{1}{2\lambda}} \left( \hat{L}^n(Q) + \lambda \left( \frac{KL(Q, P) + \ln \frac{1}{\delta}}{n} \right) \right)$$

Corollary:

$$L_{\text{probit}}(w) \leq \frac{1}{1 - \frac{1}{2\lambda_n}} \left( \hat{L}^n_{\text{probit}}(w) + \lambda_n \left( \frac{\frac{1}{2}\|w\|^2 + \ln \frac{1}{\delta}}{n} \right) \right)$$
Consistency of Ramp Loss

Now we consider the following ramp loss training equation.

\[ \hat{w}_n = \arg\min_w \hat{L}_\text{ramp}(w) + \frac{\gamma_n}{2n} \|w\|^2 \]  \hspace{1cm} (1)

If

- \( \gamma_n / (\ln^2 n) \) increases without bound
- \( \gamma_n / (n \ln n) \) converges to zero,

then

\[ \lim_{n \to \infty} L_{\text{probit}}((\ln n)\hat{w}_n) = L^* \]
Main Lemma

\[
\lim_{\sigma \to 0} L_{\text{probit}}(w/\sigma, x, y) \leq L(w, x, y) \leq L_{\text{ramp}}(w, x, y)
\]

\[
L_{\text{probit}} \left( \frac{w}{\sigma}, x, y \right) \leq L_{\text{ramp}}(w, x, y) + \sigma + \sigma \sqrt{8 \ln \frac{|Y|}{\sigma}}
\]
Proof of Main Lemma Part I

\[ L_{\text{probit}} \left( \frac{w}{\sigma}, x, y \right) \leq \sigma + \max_{\hat{y}: m(\hat{y}) \leq M} L(y, \hat{y}) \]

where

\[ m(\hat{y}) = w^\top \Delta \phi(\hat{y}) \quad \Delta \phi(\hat{y}) = \phi(x, \hat{y}_w(x)) - \phi(x, \hat{y}) \quad M = \sigma \sqrt{8 \ln \frac{|\mathcal{Y}|}{\sigma}} \]

Proof: for \( m(\hat{y}) > M \) we have the following.

\[ P_{\epsilon}[\hat{y}_{w+\sigma \epsilon}(x) = \hat{y}] \leq P_{\epsilon}[(w + \sigma \epsilon)^\top \Delta \phi(\hat{y}) \leq 0] = P_{\epsilon}[-\epsilon^\top \Delta \phi(y) \geq m(\hat{y})/\sigma] \]
\[ \leq P_{\epsilon \sim \mathcal{N}(0,1)} \left[ \epsilon \geq \frac{M}{2\sigma} \right] \leq \exp \left( -\frac{M^2}{8\sigma^2} \right) = \frac{\sigma}{|\mathcal{Y}|} \]

\[ \mathbb{E}_{\epsilon} \left[ L(y, \hat{y}_{w+\sigma \epsilon}(x)) \right] \leq P_{\epsilon} [\exists \hat{y} : m(\hat{y}) > M \, \hat{y}_{w+\epsilon \sigma}(x) = \hat{y}] + \max_{\hat{y}: m(\hat{y}) \leq M} L(y, \hat{y}) \]
\[ \leq \sigma + \max_{\hat{y}: m(\hat{y}) \leq M} L(y, \hat{y}) \]
Proof of Main Lemma Part II

\[ L_{\text{probit}} \left( \frac{w}{\sigma}, x, y \right) \leq \sigma + \max_{\hat{y}: m(\hat{y}) \leq M} L(y, \hat{y}) \]

\[ \leq \sigma + \left( \max_{\hat{y}: m(\hat{y}) \leq M} L(y, \hat{y}) - m(\hat{y}) \right) + M \]

\[ \leq \sigma + \left( \max_{\hat{y}} L(y, \hat{y}) - m(\hat{y}) \right) + M \]

\[ = \sigma + L_{\text{ramp}}(w, x, y) + M \]
Using the Main Lemma

\[ L_{\text{probit}} \left( \frac{w}{\sigma} \right) \leq \frac{1}{1 - \frac{1}{2\lambda_n}} \left( \hat{L}_n^{\text{ramp}}(w) + \sigma + \sigma \sqrt{8 \ln \frac{|Y|}{\sigma}} + \lambda_n \left( \frac{||w||^2}{2\sigma^2} + \ln \frac{1}{\delta} \right) \right) \]

Now take

\[ \sigma_n = 1 / \ln n \]

\[ \lambda_n = \gamma_n / (\ln^2 n) \]
A Comparison of Convergence Rates

Optimizing $\sigma$ as a function of $\lambda$, $||w||$ and $n$ we get (approximately).

$$\sigma = \left( \lambda_n ||w||^2 / n \right)^{1/3}$$

which gives a guarantee of

$$\frac{1}{1 - \frac{1}{2\lambda_n}} \left( L^n_{\text{ramp}}(w) + \left( \frac{\lambda_n ||w||^2}{n} \right)^{1/3} \left( \frac{3}{2} + \sqrt{8 \ln |Y| / \sigma} \right) + \frac{\lambda_n \ln \frac{1}{\delta}}{n} \right)$$

which should be contrasted with

$$L^n_{\text{probit}}(w) \leq \frac{1}{1 - \frac{1}{2\lambda_n}} \left( L^n_{\text{probit}}(w) + \lambda_n \left( \frac{1}{2} ||w||^2 + \ln \frac{1}{\delta} \right) \right)$$
Summary

- Well known Surrogate loss functions have natural generalizations to the latent structural setting.

- Convex loss functions are not consistent.

- Probit and Ramp loss are consistent but seem significantly different in the latent structural setting.