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

David McAllester
TTI-Chicago

Motivation

- Systems are often evaluated with some quantitative performance metric or *task loss*.

- Here we consider consistency in the sense of minimizing task loss. SVMs are not consistent in this sense.

- Here we show that,unlike hinge loss or log loss, latent-structural probit loss and ramp loss are both consistent.

- We also give finite-sample generalizations bounds and point to a variety of supporting empirical work.

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

$$\hat{y}_w(x) = \operatorname*{argmax}_y \; w^\top \boldsymbol{\phi}(x, y)$$

We would like

$$w^* = \operatorname*{argmin}_w \; \mathrm{E}_{x,y} \left[L(y, \hat{y}_w(x)) \right]$$

We get

$$\hat{w} = \operatorname*{argmin}_w \; \left(\sum_{i=1}^n L_s(w, x_i, y_i) \right) + \frac{\lambda}{2} ||w||^2$$

The surrogate loss L_s must be scale-sensitive and hence different form the task loss L

Standard Surrogate Loss Functions (Binary Case)

$$y \in \{-1, 1\}$$

$$\hat{y}_w(x) = \text{sign } w^\top \Phi(x) \quad \left(\text{for } \Phi(x,y) = \frac{1}{2} y \Phi(x)\right)$$

$$m = y w^\top \boldsymbol{\phi}(x)$$

$$\begin{aligned}
L_{\text{log}}(w,x,y) &= \ln\left(1 + e^{-m}\right) \\
L_{\text{hinge}}(w,x,y) &= \max(0, 1-m) \\
L_{\text{ramp}}(w,x,y) &= \min(1, \max(0, 1-m)) \\
L_{\text{probit}}(w,x,y) &= P_{\epsilon \sim \mathcal{N}(0,1)}[\epsilon \geq m] \;\; (\text{assuming } ||\Phi(x)|| = 1)
\end{aligned}$$

Latent Labels

We now assume a finite set $\mathcal{Z}$ of "latent labels".

$$\hat{s}_w(x) = \operatorname*{argmax}_{(y,z)} w^\top \boldsymbol{\phi}(x, y, z)$$

$$w^* = \operatorname*{argmin}_{w} \; \mathrm{E}_{x,y}\left[L(y, \hat{s}_w(x))\right]$$

$$L(y, (\hat{y}, \hat{z})) = L(y, \hat{y})$$

Surrogate Loss Functions

$$L_{\mathrm{log}}(w,x,y) \;=\; \ln \frac{1}{P_w(y|x)} = \ln Z_w(x) - \ln Z_w(x,y)$$

$$Z_w(x) \;=\; \textstyle\sum_{y,z} \exp(w^\top \Phi(x,y,z))$$

$$Z_w(x,y) \;=\; \textstyle\sum_{z} \exp(w^\top \boldsymbol{\phi}(x,(y,z)))$$

$$L_{\mathrm{hinge}}(w,x,y) \;=\; \left(\max_s w^\top \boldsymbol{\phi}(x,s) + L(y,s)\right) - \left(\max_z w^\top \Phi(x,(y,z))\right)$$

$$L_{\mathrm{ramp}}(w,x,y) \;=\; \left(\max_s w^\top \boldsymbol{\phi}(x,s) + L(y,s)\right) - \left(\max_s w^\top \Phi(x,s)\right)$$

$$L_{\mathrm{probit}}(w,x,y) \;=\; \mathrm{E}_\epsilon \left[L(y,\hat{s}_{w+\epsilon}(x))\right]$$

Citations

L_{log}:

CRFs: J. Lafferty, A. McCallum, and F. Pereira. 2001.

Hidden: A. Quattoni, S. Wang, L.P. Morency, M Collins, and T Darrell., 2007.

L_{hinge}:

Struct: B. Taskar, C. Guestrin, and D. Koller. 2003., I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, 2004.

Hidden: Chun-Nam John Yu and T. Joachims. 2009.

L_{ramp}:

Binary: R.Collobert, F.H.Sinz, J.Weston, and L.Bottou, 2006.

Struct: Chuong B. Do, Quoc Le, Choon Hui Teo, Olivier Chapelle, and Alex Smola, 2008.

L_{probit}: Joseph Keshet, David McAllester, and Tamir Hazan. ICASSP, 2011

Tighter Upper Bound on Task Loss

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

$L_{\text{hinge}}(w,x,y) \geq L_{\text{ramp}}(w,x,y)$ is immediate (see definitions).

Furthermore:

$$\begin{aligned} L_{\text{ramp}}(w,x,y) &= \left(\max_s w^\top \Phi(x,s) + L(y,s)\right) - w^\top \Phi(x,\hat{s}_w(x)) \\ &\geq w^\top \Phi(x,\hat{s}_w(x)) + L(y,\hat{s}_w(x)) - w^\top \Phi(x,\hat{s}_w(x)) = L(y,\hat{s}_w(x)) \end{aligned}$$

Ramp Loss (or Direct Loss) Perceptron Updates

Optimizing L_{ramp} through subgradient descent yields the following update rule (here we ignore regularization).

$$\Delta w \;\propto\; \boldsymbol{\phi}(x, \hat{s}_w(x)) - \boldsymbol{\phi}(x, \hat{s}^+_w(x, y))$$

$$\hat{s}^+_w(x, y) \;=\; \underset{s}{\mathrm{argmax}}\;\; w^\top \boldsymbol{\phi}(x, s) + L(y, s)$$

under mild conditions on the probability distribution over pairs (x, y) we have the following [McAllester, Hazan, Keshet, 2010].

$$\nabla_w L(w) = \lim_{\alpha \to \infty} \alpha \mathrm{E}_{x,y} \left[\boldsymbol{\phi}(x, \hat{s}^+_{\alpha w}(x, y)) - \boldsymbol{\phi}(x, \hat{s}_{\alpha w}(x)) \right]$$

Empirical Results

Methods closely related to ramp loss updates with early stopping have been shown to be effective in machine translation.

- P. Liang, A. Bouchard-Ct, D. Klein, and B. Taskar. (COLING/ACL), 2006.
- D. Chiang, K. Knight, and W. Wang. NAACL, 2009

Ramp loss updates regularized with early stopping (direct loss) has been shown to give improvements over hinge loss in phoneme alignment and phoneme recognition.

- McAllester, Hazan, and Keshet. NIPS 2010.
- Keshet, Cheng, Stoehr, and McAllester, to appear at Interspeech 2011

Probit loss has been show to give an improvement over hings loss for phoneme recognition.

- Keshet, McAllester, and Hazan, ICASSP, 2011.

Some Notation

$$L(w) = \mathrm{E}_{x,y}\left[L(w, x, y)\right]$$

$$L^* = \inf_w L(w)$$

$$\hat{L}^n(w) = \frac{1}{n}\sum_{i=1}^{n} L(w, x_i, y_i)$$

Consistency of Probit Loss

We consider the following learning rule where λ_n is some given function of n.

$$\hat{w}_n = \operatorname{argmin} w \;\; \hat{L}^n_{\text{probit}}(w) \;\; + \;\; \frac{\lambda_n}{2n}\, ||w||^2$$

If

- λ_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 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.

$$\hat{w}_n = \mathrm{argmin}\, w \;\; \hat{L}^n_{\mathrm{ramp}}(w) \;\; + \;\; \frac{\gamma_n}{2n}\, ||w||^2 \tag{1}$$

If

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

then

$$\lim_{n\to\infty} L_{\mathrm{probit}}((\ln n)\hat{w}_n) = L^*$$

Main Lemma

$$\lim_{\sigma \to 0} L_{\mathrm{probit}}(w/\sigma, x, y) \leq L(w, x, y) \leq L_{\mathrm{ramp}}(w, x, y)$$

.

$$L_{\mathrm{probit}}\left(\frac{w}{\sigma}, x, y\right) \leq L_{\mathrm{ramp}}(w, x, y) + \sigma + \sigma\sqrt{8 \ln \frac{|\mathcal{S}|}{\sigma}}$$

Proof of Main Lemma Part I

$$L_{\text{probit}}\left(\frac{w}{\sigma}, x, y\right) \leq \sigma + \max_{s\colon m(s)\leq M} L(y, s)$$

where

$$m(s) = w^\top \Delta\boldsymbol{\phi}(s) \qquad \Delta\boldsymbol{\phi}(s) = \boldsymbol{\phi}(x,\ \hat{s}_w(x)) - \boldsymbol{\phi}(x,\ s) \qquad M = \sigma\sqrt{8\ln\frac{|\mathcal{S}|}{\sigma}}$$

Proof: for $m(s) > M$ we have the following.

$$\begin{aligned} P_\epsilon[\hat{s}_{w+\sigma\epsilon}(x) = s] \ &\leq\ P_\epsilon[(w+\sigma\epsilon)^\top \Delta\boldsymbol{\phi}(s) \leq 0] = P_\epsilon\left[-\epsilon^\top\Delta\boldsymbol{\phi}(s) \geq m(s)/\sigma\right] \\ &\leq\ P_{\epsilon\sim\mathcal{N}(0,1)}\left[\epsilon \geq \frac{M}{2\sigma}\right] \leq \exp\left(-\frac{M^2}{8\sigma^2}\right) = \frac{\sigma}{|\mathcal{S}|} \end{aligned}$$

$$\begin{aligned} \mathrm{E}_\epsilon\left[L(y, \hat{s}_{w+\sigma\epsilon}(x))\right] \ &\leq\ P_\epsilon\left[\exists s :\ m(s) > M \ \ \hat{s}_{w+\epsilon\sigma}(x) = s\right] + \max_{s:m(s)\leq M} L(y, s) \\ &\leq\ \sigma + \max_{s:m(s)\leq M} L(y, s) \end{aligned}$$

Proof of Main Lemma Part II

$$
\begin{aligned}
L_{\text{probit}}\left(\frac{w}{\sigma}, x, y\right) &\leq \sigma + \max_{s\colon m(s) \leq M} L(y, s) \\
&\leq \sigma + \left(\max_{s\colon m(s) \leq M} L(y, s) - m(s)\right) + M \\
&\leq \sigma + \left(\max_{s} L(y, s) - m(s)\right) + M \\
&= \sigma + L_{\text{ramp}}(w, x, y) + M
\end{aligned}
$$

Using the Main Lemma

$$L_{\mathrm{probit}}\left(\frac{w}{\sigma}\right) \leq \frac{1}{1-\frac{1}{2\lambda_n}}\left(\hat{L}^n_{\mathrm{ramp}}(w)+\sigma+\sigma\sqrt{8\ln\frac{|\mathcal{S}|}{\sigma}}+\lambda_n\left(\frac{\frac{||w||^2}{2\sigma^2}+\ln\frac{1}{\delta}}{n}\right)\right)$$

Now take

$$\sigma_n = 1/\ln n$$

$$\lambda_n = \gamma_n/(\ln^2 n)$$

A Comparison of Convergence Rates

Optimizing σ as a function of λ, $||w||$ and n we get (approximately).

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

which gives

$$L_{\mathrm{probit}}\left(\frac{w}{\sigma}\right) \leq \frac{1}{1-\frac{1}{2\lambda_n}}\left(\hat{L}^n_{\mathrm{ramp}}(w) + \left(\frac{\lambda_n ||w||^2}{n}\right)^{1/3}\left(\frac{3}{2} + \sqrt{8 \ln \frac{|\mathcal{S}|}{\sigma}}\right) + \frac{\lambda_n \ln \frac{1}{\delta}}{n}\right)$$

which should be contrasted with

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

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.

Future Work

Early empirical evidence suggests that “direct loss” early stopping (without other regularization) works significantly better than L_2 regularization for ramp loss.

The direct loss theorem holds for the “toward better” variant of ramp loss. This variant works better in practice.

This provides both empirical and theoretical evidence that the analysis presented here is too weak.