Some PAC-Bayesian Theorems

David McAllester
TTI-Chicago

SVM

$$w^* = \underset{w}{\operatorname{argmin}} \sum_i \max\left(0, 1 - y_i w^T \Phi(x_i)\right) + \frac{1}{2}\lambda ||w||^2$$

For $||\Phi(x)|| = 1$ SVMlight default is equivalent to $\lambda = 1$.

The default $\lambda = 1$ "holds up" independent of n and independent of the number of support vectors.

Why?

Bayesian Inference

$$h^* = \operatorname*{argmin}_h \sum_i \ln \frac{1}{P(y_i|x_i;h)} + \ln \frac{1}{P(h)}$$

Note that $\lambda = 1$.

Failure of Square Root Bounds

$$\mathrm{err}(w) \;\leq\; \frac{1}{n}\sum_i I\left[y_i w^T \Phi(x_i) \leq 1\right] + O\left(\sqrt{\frac{||w||^2}{n}}\right)$$

$$n\;\mathrm{err}(w) \;\leq\; \sum_i I\left[y_i w^T \Phi(x_i) \leq 1\right] + O\left(\sqrt{n}\;||w||\right)$$

$$||w_0 + \Delta w|| \;\approx\; ||w_0|| + \frac{(\Delta w)^T w_0}{||w_0||} \approx a + \frac{||w_0 + \Delta w||^2}{2||w_0||}$$

$$\lambda \;\approx\; O\left(\sqrt{\frac{n}{||w_0||^2}}\right)$$

PAC-Bayesian Theorem

$$\mathrm{err}(Q) \;\leq\; B(Q)$$

$$B(Q) \;\doteq\; \widehat{\mathrm{err}}(Q) + \sqrt{\widehat{\mathrm{err}}(Q)c(Q)} + c(Q)$$

$$c(Q) \;\doteq\; \frac{2\left(KL(Q,P) + \ln\frac{n+1}{\delta}\right)}{n}$$

$$\widehat{\mathrm{err}}(Q) + c(Q) \;\leq\; B(Q) \quad \leq \frac{3}{2}\left(\widehat{\mathrm{err}}(Q) + c(Q)\right)$$

$$Q^* \;\approx\; \operatorname*{argmin}_Q \; \sum_i \mathrm{E}_{w\sim Q}\left[L(w,x_i,y_i)\right] + 2KL(Q,P)$$

This provides a rationalization of $\lambda = 1$.

L_2 **Prior**

[Langford,Shawe-Taylor 2002, McAllester 2003]

Prior $$P(w) = \frac{1}{Z}\, e^{-\frac{||w||^2}{2\sigma^2}}$$

Posterior $$Q_\mu(w) = \frac{1}{Z} e^{-\frac{||w-\mu||^2}{2\sigma^2}}$$

$$\widehat{\text{err}}(Q) = \frac{1}{n} \sum_{i=1}^{n} L_{\text{probit}} \left(\frac{y_i \mu^T \Phi(x_i)}{\sigma ||\Phi(x_i)||_2} \right)$$

$$L_{\text{probit}}(z) = P_{u \sim \mathcal{N}(0,1)}[u \geq z]$$

$$KL(Q_\mu, P) = \frac{||\mu||^2}{2\sigma^2}$$

$\sigma = \frac{1}{||\Phi||_2}$ $$\mu^* \approx \text{argmin}_\mu \sum_i L_{\text{probit}} \left(y_i \mu^T \Phi(x_i) \right) + ||\Phi||_2^2 ||\mu||_2^2$$

L_1 **Prior**

Prior $$P(w) = \frac{1}{Z} e^{-\frac{||w||_1}{\gamma}}$$

Posterior $$Q_\mu(w) = \frac{1}{Z} e^{-\frac{||w_i - \mu||_1}{\gamma}}$$

$$\widehat{\text{err}}(Q) \approx \frac{1}{n} \sum_{i=1}^{n} L_{\text{probit}} \left(\frac{y_i \mu^T \Phi(x_i)}{2\gamma ||\Phi(x)||_2} \right)$$

$$KL(Q_\mu, P) \approx \frac{||\mu||_1}{\gamma}$$

$\gamma = \frac{1}{2||\Phi||_2}$ $$\mu^* \approx \text{argmin}_\mu \sum_i L_{\text{probit}} \left(y_i \mu^T \Phi(x_i) \right) + 4||\Phi||_2 ||\mu||_1$$

L_0 Prior

[Schapire,Freund,Bartlett,Lee,98] [Langford, Seeger, Meggiddo, 2001]

Prior	$P(w)$	N independent feature draws (uniform)
Posterior	$Q_\mu(w)$	N independent feature draws from $\frac{\mu}{\|\|\mu\|\|_1}$

$$\widehat{\mathrm{err}}(Q) \approx \frac{1}{n}\sum_{i=1}^{n} L_{\mathrm{probit}}\left(\frac{y_i \mu^T \Phi(x_i)}{\frac{1}{2\sqrt{N}}||\Phi(x)||_\infty ||\mu||_1}\right)$$

$$KL(Q_\mu, P) \approx N \ln d$$

$$||\mu||_1 = \frac{2\sqrt{N}}{||\Phi||_\infty} \qquad \mu^* \approx \operatorname{argmin}_\mu \sum_i L_{\mathrm{probit}}\left(y_i \mu^T \Phi(x_i)\right) + \frac{1}{2}||\Phi||_\infty^2 (\ln d) ||\mu||_1^2$$

The Hinge Loss Problem

For $||\Phi(x)||_2 = 1$, is there an approximation guarantee between

$$w^* = \underset{w}{\operatorname{argmin}} \sum_i L_{\text{probit}}(y_i w^T \Phi(x_i)) + ||w||^2$$

and

$$w^* = \underset{w}{\operatorname{argmin}} \quad \sum_i \max\left(0, 1 - y_i w^T \Phi(x_i)\right) + ||w||^2$$

Structured Prediction with an L_2 Prior

Consider machine translation where x is an English sentence and y is a French sentence.

$$y_w(x) = \underset{y}{\operatorname{argmax}} \; w^T \Phi(x, y)$$

Consider a loss function $L(y, y_w(x))$ such as the BLEU score.

$$P(w) = \frac{1}{Z} e^{-\frac{||w||^2}{2\sigma^2}}$$

$$Q_\mu(w) = \frac{1}{Z} e^{-\frac{||w-\mu||^2}{2\sigma^2}}$$

$$\mu^* \approx \underset{\mu}{\operatorname{argmin}} \; \sum_i \mathrm{E}_{w \sim Q_\mu} [L(y_i, y_w(x_i))] + \frac{1}{\sigma^2} ||\mu||_2^2$$

Digression: Ignore Regularization

$$w^* = argmin_w \sum_i L(y_i, y_w(x_i))$$

Many authors work with the following convex relaxation — the so-called structured hinge loss.

$$\mathrm{margin}_i(\hat{y}) \doteq w^T \Phi(x_i, y_i) - w^T \Phi(x_i, \hat{y})$$

$$\begin{aligned} L(y_i, y_w(x_i)) &\leq L(y_i, y_w(x_i)) - \mathrm{margin}_i(y_w(x_i)) \\ &\leq \max_{\hat{y}} L(y_i, \hat{y}) - \mathrm{margin}_i(\hat{y}) \\ &= \max(0, 1 - y_i w^T \Phi(x_i)) \;\; \text{for } y \in \{-1, 1\} \end{aligned}$$

Structured Hinge Generalizes Binary Hinge

Under Hamming loss, Grouping binary training data into bags and applying structured hinge to each bag is equivalent to binary hinge on the original data.

$$\text{structured hinge:}\quad w^* = \mathrm{argmin}_w \left(\sum_i \max_y H(y_i, y) - m_i(y) \right) + \frac{1}{2}\lambda ||w||^2$$

$$\text{binary hinge:}\quad w^* = \mathrm{argmin}_w \sum_i \max_{y \in \{-1,1\}} I[y \neq y_i] - m_i(y) + \frac{1}{2}\lambda ||w||^2$$

Margin Bounds

$$n \; \mathrm{E}_{(x,y)\sim\rho}\left[L(y, y_w(x))\right] \leq O\left(\sum_i \max_{y:\ \mathrm{margin}_i(y) \leq H(y,y_i)} L(y_i, y) \quad + \quad ||w||^2\right)$$

This involves both the Hamming distance (as a margin requirement) and the loss function.

Perceptron-like Updates

For a training point (x, y) we consider:

multiclass perceptron: $\Delta w \propto \Phi(x, y) - \Phi(x, \hat{y})$

structured hinge subgradient: $\Delta w \propto \Phi(x, y) - \Phi(x, \hat{y}_{\mathrm{hinge}})$

$$\hat{y} = \operatorname*{argmax}_{\hat{y}} \; w^T \Phi(x, \hat{y})$$

$$\hat{y}_{\mathrm{hinge}} = \operatorname*{argmax}_{\hat{y}} \; w^T \Phi(x, \hat{y}) + L(y, \hat{y})$$

The optimization problem defining $\hat{y}_{\mathrm{hinge}}$ is called loss adjusted inference.

Direct Loss Update

Joint work with Tamir Hazan and Joseph Keshet

For a training point (x, y) we consider:

$$\text{direct loss:} \quad \Delta w \;\propto\; \Phi(x, \hat{y}_L) - \Phi(x, \hat{y})$$

$$\hat{y} \;=\; \operatorname*{argmax}_{\hat{y}} \;\; w^T\Phi(x, \hat{y})$$

$$\hat{y}_L \;=\; \operatorname*{argmax}_{\hat{y}} \;\; w^T\Phi(x, \hat{y}) - \epsilon L(y, \hat{y})$$

Updates similar to the loss minimization appear in
[Liang, Bouchard-Côté, Klein, and Taskar, 2006] [Chiang,Knight,Wang,2009]

Direct Loss Theorem

If, for each u, we have $p(\Phi(x,u)|y=u)$ is a continuous density on $\mathbb{R}^d$ then we have the following.

$$-\nabla_w \mathrm{E}_{(x,y)\sim\rho}\left[L(y,\hat{y})\right] = \lim_{\epsilon\to 0} \frac{\mathrm{E}_{(x,y)\sim\rho}\left[\Phi(x,\hat{y}_L) - \Phi(x,\hat{y})\right]}{\epsilon}$$

$$\hat{y} = \operatorname*{argmax}_{\hat{y}} \; w^T\Phi(x,\hat{y})$$

$$\hat{y}_L = \operatorname*{argmax}_{\hat{y}} \; w^T\Phi(x,\hat{y}) - \epsilon L(y,\hat{y})$$

Proof Hint

$$\left(\nabla_w \mathrm{E}_{(x,y)\sim\rho}\left[L(y,\hat{y}(w))\right]\right)^T \Delta w$$

$$= \sum_{u,v} \lim_{\epsilon\to 0} \frac{1}{\epsilon} \quad \mathrm{E}_{(x,y)\sim\rho}\left[I\left[w^T\Delta\Phi_{u,v}(x) \in \left(0, \epsilon(\Delta w)^T\Delta\Phi_{v,u}(x)\right)\right] \quad \Delta L_{v,u}(y)\right]$$

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon} \quad \mathrm{E}_{(x,y)\sim\rho}\left[\Phi(x,\hat{y}_L) - \Phi(x,\hat{y})\right]$$

$$= \sum_{u,v} \lim_{\epsilon\to 0} \frac{1}{\epsilon} \quad \mathrm{E}_{(x,y)\sim\rho}\left[I\left[w^T\Delta\Phi_{u,v}(x) \in (0, \epsilon\Delta L_{v,u}(y))\right] \quad \Delta\Phi_{v,u}(x)\right]$$

$$= \sum_{u,v} \lim_{\epsilon\to 0} \frac{1}{2\epsilon} \quad \mathrm{E}_{(x,y)\sim\rho}\left[I\left[w^T\Delta\Phi_{u,v}(x) \in (0, \epsilon)\right] \; \Delta\Phi_{v,u}(x)\; \Delta L_{v,u}(y)\right]$$

$$\Delta\Phi_{u,v}(x) \doteq \Phi(x,u) - \Phi(x,v)$$

$$\Delta L_{u,v}(y) \doteq L(y,u) - L(y,v)$$

Approximate Inference and Hidden Information

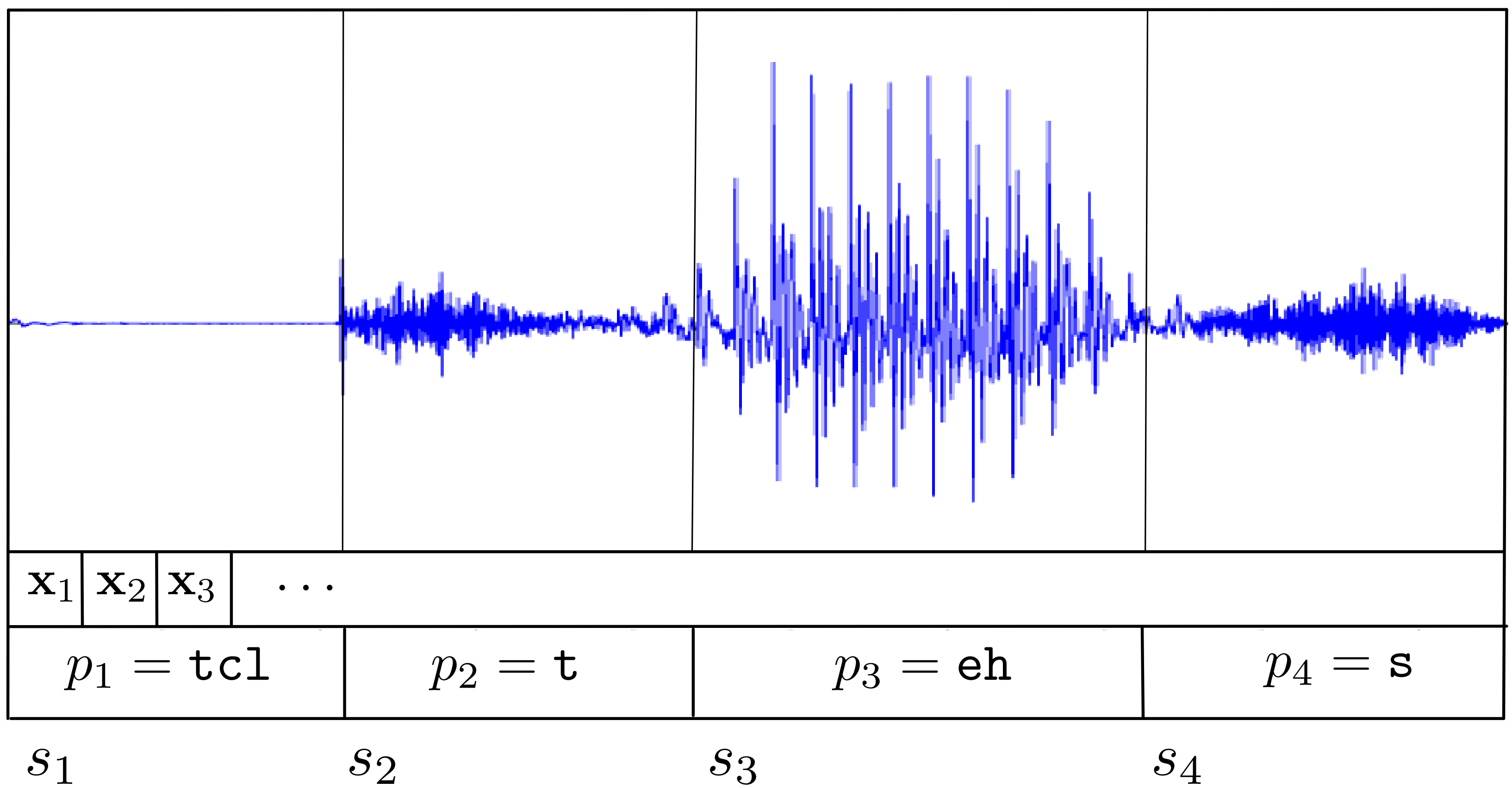
Let $\mathcal{P}$ be any finite set.

- $\mathcal{P}$ might be the set of corners of a relaxation of the marginal polytope.
- $\mathcal{P}$ might be the set of pairs $\langle y, h \rangle$ where y is a label and h is a hidden label.

For $\mu \in \mathcal{P}$ we let $\Phi(x, \mu)$ be a feature vector and $L(y, \mu)$ be a loss.

$$
\begin{aligned}
-\nabla_w \mathrm{E}_{(x,y)\sim\rho}\left[L(y, \hat{\mu}(x))\right] &= \lim_{\epsilon \to 0} \frac{\mathrm{E}_{(x,y)\sim\rho}\left[\Phi(x, \hat{\mu}_L(x, y)) - \Phi(x, \hat{\mu}(x))\right]}{\epsilon} \\
\hat{\mu}(x) &= \operatorname*{argmax}_{\hat{\mu}} \;\; w^T \Phi(x, \hat{\mu}) \\
\hat{\mu}_L(x, y) &= \operatorname*{argmax}_{\hat{\mu}} \;\; w^T \Phi(x, \hat{\mu}) - \epsilon L(y,\; \hat{\mu})
\end{aligned}
$$

Experiments (Joseph Keshet)



A spoken utterance labeled with the sequence of phonemes $/p_1\ p_2\ p_3\ p_4/$ and its corresponding sequence of start-times $(s_1\ s_2\ s_3\ s_4)$.

Loss Functions

Two types of loss functions are used in this problem:

The τ-*alignment loss*

$$L^{\tau\text{-alignment}}(\bar{s}, \bar{s}') = \frac{1}{|\bar{s}|} \left|\{i : |s_i - s'_i| > \tau\}\right|$$

The τ-*insensitive loss*

$$L^{\tau\text{-insensitive}}(\bar{s}, \bar{s}') = \frac{1}{|\bar{s}|} \max\left\{|s_i - s'_i| - \tau, 0\right\}$$

Results

TIMIT code test set (192 utterances):

	$\tau \leq 10$ms	$\tau \leq 20$ms	$\tau \leq 30$ms	$\tau \leq 40$ms
Keshet *et al* (2007)	79.7	92.1	96.2	98.1
Direct Loss Min τ-alignment loss	85.83	94.05	97.04	98.17
Direct Loss Min τ-insensitive loss	**86.00**	**94.48**	**97.20**	**98.47**

TIMIT the whole test-set (1344 utterances).

	$\tau \leq 10$ms	$\tau \leq 20$ms	$\tau \leq 30$ms	$\tau \leq 40$ms
Hosom (2009)	79.30	93.36	96.74	98.22
Keshet *et al* (2007)	80.0	92.3	96.4	98.2
Direct Loss Min τ-alignment loss	**86.01**	94.08	97.08	98.44
Direct Loss Min τ-insensitive loss	85.72	**94.21**	**97.21**	**98.60**

Differentiating the PAC-Bayes bound

Now differentiate the PAC-Bayes bound with respect to μ under L_2 regularization.

$$
\begin{aligned}
&\nabla_\mu \left(\mathrm{E}_{w \sim Q_\mu} \left[\sum_i L(y_i, \hat{y}(w, x_i)) \right] \right) \\
&= \sum_{i=1}^{n} \nabla_\mu \left(\int Q_\mu(w) L(y_i, \hat{y}(w, x_i)) dw \right) \\
&= \sum_{i=1}^{n} \int Q_\mu(w)(w - \mu) L(y_i, \hat{y}(w, x_i)) dw \\
&= \sum_{i=1}^{n} \mathrm{E}_{w \sim Q_\mu} \left[(w - \mu) L(y_i, \hat{y}(w, x_i)) \right] \\
&= \sum_{i=1}^{n} \frac{1}{2} \mathrm{E}_{\Delta w \sim P} \left[\Delta w \left(L(y_i, \hat{y}(w + \Delta w, x_i)) - L(y_i, \hat{y}(w - \Delta w, x_i)) \right) \right]
\end{aligned}
$$

Summary

- PAC-Bayesian bounds predict λ (the regularization paraeter).
- PAC-Bayesian bounds allow L_2, L_1 and L_0 regularization to be understood in terms of prior probabilities.
- Hinge loss — both binary and structured — is a convex relaxation which has no know approximation guarantee.
- Existing theories of structured learning confuse margin requirements with loss functions.
- Direct loss optimiztion, or bound optimization, is an up and coming approach to structured learning.