Supplement: Marginal Structured SVM with Hidden Variables

Wei Ping WPING@ICS.UCI.EDU

Qiang Liu QLIU1@ICS.UCI.EDU

Alexander Ihler IHLER@ICS.UCI.EDU

Department of Computer Science, UC Irvine

Constraint Form of Marginal Structured SVM

Here we give the constraint form of Eq. (3) in the main paper,

$$\min_{w,\{\xi_i\geq 0\}} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i, \tag{1}$$
$$\text{s.t. } \forall i \in \{1,...,n\},\ \forall y \in \mathcal{Y},\ \log\sum_h \exp\Big[w^T\phi(x_i,y_i,h)\Big] - \log\sum_h \exp[w^T\phi(x_i,y,h)] \geq \Delta(y_i,y) - \xi_i,$$

where $\{\xi_i\}$ are the slack variables. One can show that the optimal solution $\{\xi_i^*\}, w^*$ satisfies,

$$\xi_i^* = \max_y \Big\{\Delta(y_i,y) + \log\sum_h \exp[w^{*T}\phi(x_i,y,h)]\Big\} - \log\sum_h \exp[w^{*T}\phi(x_i,y_i,h)],$$

which gives the same objective value as the the unconstrained form. One can also derive a cutting plane-based training algorithm (Joachims et al., 2009) for this constraint formulation.

Details of Proofs

In this section, we give proofs for two lemmas referenced but omitted from the main paper.

Lemma 1. *The objective of the unified framework (Eq. (5) in main paper) is an upper bound of the empirical loss function* $\Delta(y_i, \hat{y_i}^{\epsilon_h}(w))$ *over the training set, where the prediction* $\hat{y_i}^{\epsilon_h}(w)$ *is decoded by "annealed" marginal MAP,*

$$\hat{y_i}^{\epsilon_h}(w) = \arg\max_y \log\sum_h \exp\Big[\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big].$$

Proof.

$$\begin{aligned}\Delta(y_i,\hat{y_i}^{\epsilon_h}(w)) &\leq \Delta(y_i,\hat{y_i}^{\epsilon_h}(w)) + \epsilon_h\log\sum_h \exp\big[\frac{w^T\phi(x_i,\hat{y_i}^{\epsilon_h}(w),h)}{\epsilon_h}\big] - \epsilon_h\log\sum_h\exp\big[\frac{w^T\phi(x_i,y_i,h)}{\epsilon_h}\big]\\ &\leq \epsilon_y\log\sum_y\exp\Big\{\frac{1}{\epsilon_y}\Big[\Delta(y_i,y)+\epsilon_h\log\sum_h\exp\Big(\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big)\Big]\Big\} - \epsilon_h\log\sum_h\exp\big[\frac{w^T\phi(x_i,y_i,h)}{\epsilon_h}\big],\end{aligned}$$

where the first inequality holds by the definition of $\hat{y_i}^{\epsilon_h}(w)$, and the second holds for $\forall\epsilon_y > 0$, because the summation over y contains $\hat{y_i}^{\epsilon_h}(w)$. □

For convenience, we denote this upper bound as

$$U_i(w;\epsilon_y,\epsilon_h) = U_i^+(w;\epsilon_y,\epsilon_h) - U_i^-(w;\epsilon_h) \tag{2}$$

where

$$U_i^+(w;\epsilon_y,\epsilon_h) = \epsilon_y \log \sum_y \exp\Big\{\frac{1}{\epsilon_y}\Big[\Delta(y_i,y) + \epsilon_h \log \sum_h \exp\Big(\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big)\Big]\Big\}$$

$$U_i^-(w;\epsilon_h) = \epsilon_h \log \sum_h \exp\big[\frac{w^T\phi(x_i,y_i,h)}{\epsilon_h}\big].$$

Lemma 2. *The (sub-)gradient of $U_i(w;\epsilon_y,\epsilon_h)$ in (2) is,*

$$\nabla_w U_i(w;\epsilon_y,\epsilon_h) = \mathbb{E}_{p^{(\epsilon_y,\epsilon_h)}(y,h|x_i)}[\phi(x_i,y,h)] - \mathbb{E}_{p^{\epsilon_h}(h|x_i,y_i)}[\phi(x_i,y_i,h)],$$

where the corresponding temperature controlled distribution is defined as,

$$p^{\epsilon_h}(h|x_i,y) \propto \exp\Big[\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big],$$

$$p^{(\epsilon_y,\epsilon_h)}(y|x_i) \propto \exp\Big\{\frac{1}{\epsilon_y}\big[\Delta(y,y_i) + \epsilon_h \log \sum_h \exp\big(\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\big)\big]\Big\},$$

$$p^{(\epsilon_y,\epsilon_h)}(y,h|x_i) = p^{\epsilon_h}(h|x_i,y)\cdot p^{(\epsilon_y,\epsilon_h)}(y|x_i).$$

Proof.

$$\begin{aligned}
\nabla_w\Big(\epsilon_h \log \sum_h \exp\big[\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\big]\Big) &= \epsilon_h \frac{\sum_h \Big\{\exp\Big[\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big]\cdot\Big[\frac{\phi(x_i,y,h)}{\epsilon_h}\Big]\Big\}}{\sum_h \exp\Big[\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big]} \\
&= \sum_h \Big\{\frac{\exp\Big[\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big]}{\sum_h \exp\Big[\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big]}\cdot\phi(x_i,y,h)\Big\} \\
&= \mathbb{E}_{p^{\epsilon_h}(h|x_i,y)}[\phi(x_i,y,h)] \qquad (3)
\end{aligned}$$

As a result, $\nabla_w U_i^-(w;\epsilon_h) = \mathbb{E}_{p^{\epsilon_h}(h|x_i,y_i)}[\phi(x_i,y_i,h)]$, and

$$\nabla_w U_i^+(w;\epsilon_y,\epsilon_h) = \epsilon_y \frac{\sum_y \Big\{\exp\Big\{\frac{1}{\epsilon_y}\Big[\Delta(y_i,y) + \epsilon_h \log\sum_h \exp\Big(\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big)\Big]\Big\}\cdot\frac{1}{\epsilon_y}\cdot\nabla_w\Big(\epsilon_h\log\sum_h\exp\big[\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\big]\Big)\Big\}}{\sum_y \exp\Big\{\frac{1}{\epsilon_y}\Big[\Delta(y_i,y) + \epsilon_h \log\sum_h \exp\Big(\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big)\Big]\Big\}}$$

Subinstitute the gradient result (3),

$$\begin{aligned}
&= \frac{\sum_y \Big\{\exp\Big\{\frac{1}{\epsilon_y}\Big[\Delta(y_i,y) + \epsilon_h \log\sum_h \exp\Big(\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big)\Big]\Big\}\cdot\mathbb{E}_{p^{\epsilon_h}(h|x_i,y)}[\phi(x_i,y,h)]\Big\}}{\sum_y \exp\Big\{\frac{1}{\epsilon_y}\Big[\Delta(y_i,y) + \epsilon_h \log\sum_h \exp\Big(\frac{w^T\phi(x_i,y,h)}{\epsilon_h}\Big)\Big]\Big\}} \\
&= \mathbb{E}_{p^{(\epsilon_y,\epsilon_h)}(y|x_i)}\mathbb{E}_{p^{\epsilon_h}(h|x_i,y)}[\phi(x_i,y,h)] \\
&= \mathbb{E}_{p^{(\epsilon_y,\epsilon_h)}(y,h|x_i)}[\phi(x_i,y,h)] \qquad (4)
\end{aligned}$$

which completes the proof. □

Likelihood vs. Prediction Accuracy

In our main paper, we demonstrate that our proposed MSSVM consistently outperforms HCRF on prediction accuracy. However, it is worth noting that the HCRF model always achieves higher test likelihood than the MSSVM and LSSVM on our simulated data set. As an example, Figure 1 shows the test log-likelikelihood across the different methods on these data. This should not be surprising, since the HCRF model directly optimizes the likelihood objective, and (in this case) the model class being optimized is correct (i.e., the data were drawn from a true model with the same structure). However, higher likelihood does not necessarily imply that the HCRF will have better predictions on the target variables. As was illustrated in the main paper (see details in Section 7.1, Training Sample Size), explicitly minimizing the empirical loss can lead to better predictions in situations with high dimensional model parameters and relatively few training instances.

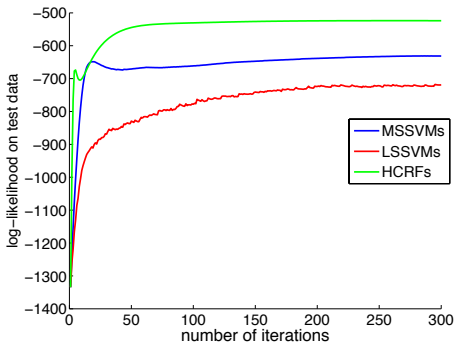

Figure 1. The test log-likelihood of MSSVM, LSSVM and HCRF using SGD when 20 training and 100 test instances are sampled from 40-node hidden chain MRF (same setting as Table 2 in main paper).

References

Joachims, T., Finley, T., and Yu, C.N.J. Cutting-plane training of structural SVMs. *Machine Learning*, 77:27–59, 2009.