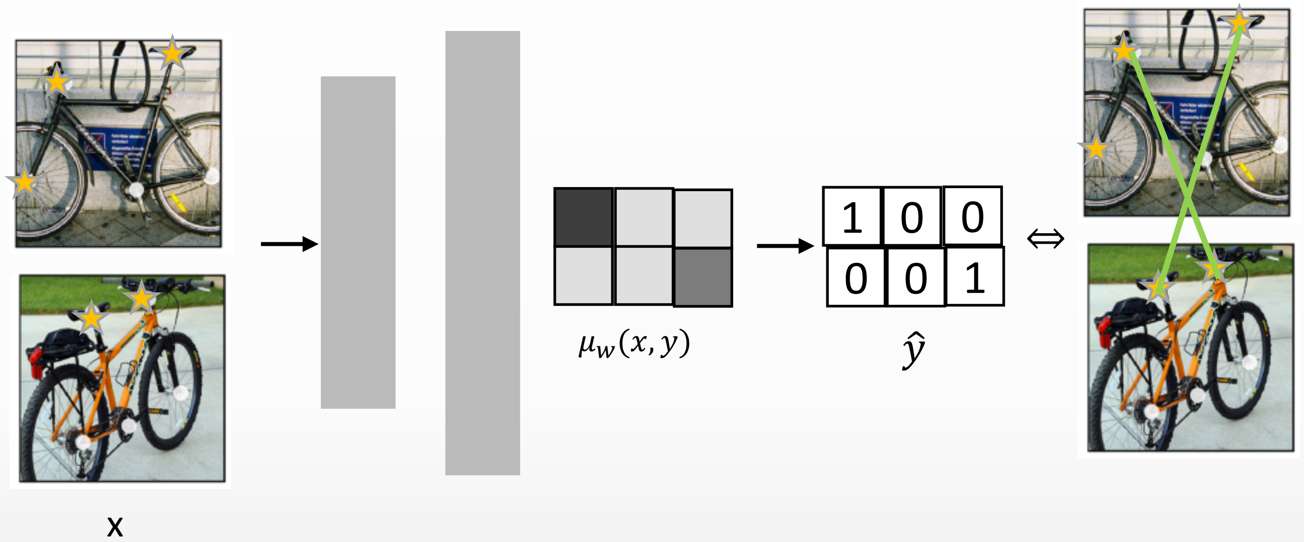
Learning Randomly Perturbed Structured Predictors for Direct Loss Minimization

Hedda Cohen Indelman, Tamir Hazan
The Technion

Motivation

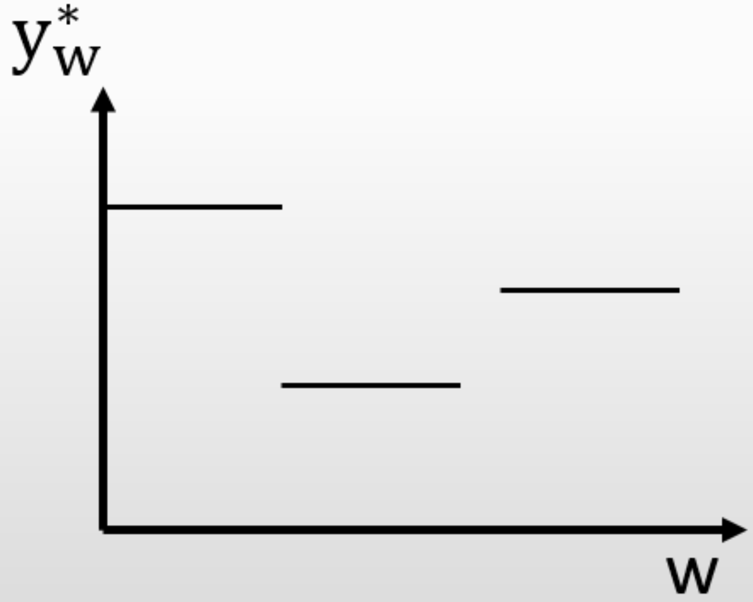
Motivation



- Learn to predict structured labels $y \in Y$ (matchings, permutations etc.) of data instances $\mathrm{x} \in X$.
- The parameters of a scoring function $\mu_w(x, y)$ are fitted to minimize the loss $\ell(y, y_w^*)$ between the label y and the highest scoring structure

Challenges in discrete labels

- The maximal argument of $\mu_w(x, y)$ is a piecewise constant function of w, and its gradient with respect to w is zero for almost any w.



Direct loss minimization

Direct loss minimization

- Let y_w^* be the highest scoring structure

$$y_w^* = \arg\max_{\hat{y} \in Y} \{\mu_w(x, \hat{y})\}$$

Direct loss minimization (Hazan et al., 2010) aims at minimizing the expected loss:

$$\min_w \mathrm{E}_{(x,y)\sim D}\, \ell(y_w^*, y)$$

Direct loss minimization

- A loss-perturbed predictor $y_w^*(\epsilon)$ is introduced:

$$y_w^*(\epsilon) = \arg\max_{\hat{y}\in Y} \{\mu_w(x,\hat{y}) + \epsilon\ell(\hat{y},y)\}$$

and the corresponding gradient estimator takes the following form:

$$\nabla_w E[\ell(y,y_w^*)] = \lim_{\epsilon\to 0}\frac{1}{\epsilon}E_{(x,y)\sim D}\left[\nabla_w\mu_w(x,y_w^*(\epsilon)) - \nabla_w\mu_w(x,y_w^*)\right]$$

- When $\epsilon < 0$, $y_w^*(\epsilon)$ returns the label with a lower loss and the gradient resembles a "moving towards better" step.
- When $\epsilon > 0$, $y_w^*(\epsilon)$ returns the label with a higher loss and the gradient resembles a "moving away from bad" step.

Direct loss minimization

1. A "general position assumption" was defined so that $w \neq 0$.

We identify that the underlying requirement is that the maximizing structure is unique.

2. It assumes smoothness of the data distribution D

Injecting noise

- Adding smooth random noise $\gamma(y)$ to $\mu_w(x, y)$ induces a probability distribution over structures y.
- [Lorberbom et al. 2018] The corresponding gradient estimator in discriminative learning setting, takes the form:

$$y^*_{w,\gamma} = \arg\max_{\hat{y} \in Y} \{\mu_w(x, \hat{y}) + \gamma(\hat{y})\}$$

$$y^*_{w,\gamma}(\epsilon) = \arg\max_{\hat{y} \in Y} \{\mu_w(x, \hat{y}) + \gamma(\hat{y}) + \epsilon\ell(\hat{y}, y)\}$$

$$\nabla_w E_\gamma\left[\ell\left(y, y^*_{w,\gamma}\right)\right] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} E_{\gamma \sim g}\left[\nabla_w \mu_w(x, y^*_{w,\gamma}(\epsilon)) - \nabla_w \mu_w\left(x, y^*_{w,\gamma}\right)\right]$$

Our contributions

Noise variance in direct loss minimization

- The random perturbation that smooths the objective might also serve as noise that masks the signal $\mu_w(x, y\,)$. To address this caveat, we learn its variance.
- By reparametrization:

$$y^*_{w,\gamma,\boldsymbol{v}} = \arg\max_{\hat{y} \in Y} \{\mu_w(x, \hat{y}) + \boldsymbol{\sigma}_{\boldsymbol{v}}(\boldsymbol{x})\gamma(\hat{y})\}$$

Connection to temperature Gumbel-max trick

- We prove that when $\gamma_i(y_i)$ are i.i.d. random variables sampled from the standard Gumbel distribution, $y^*_{w,\gamma}$ is distributed according to the Gibbs distribution, defined by the **signal-to-noise** ratio:

$$P_{\gamma\sim g}\left[\arg\max_{\hat{y}\in Y}\ \{\mu_w(x,\hat{y})+\sigma(x)\gamma(\hat{y})\} \quad = y\right] \propto e^{\boldsymbol{\mu_w(x,y)/\sigma(x)}}$$

- Thus, we make the connection between $\sigma(x)$ and temperature t in Gumbel-Softmax models.

Extending for the high-dimensional set-up

- In high-dimensional structured prediction, the number of possible structures is exponential in n.
- Scoring and sampling a noise random variable for each possible structure might be computationally intractable .

Integrating noise variance learning in direct loss minimization theorem

- We aim to learn the balance between the mean score of the randomized predictor, namely $\sum_{\alpha \in A} \mu_{w,\alpha}(x, y_\alpha)$, and the variance of its noise $\sum_{i=1}^{n} \gamma_i(\hat{y}_i)$.
- We reparameterize the randomized predictor:

$$y^*_{w,\gamma,\boldsymbol{v}} \in \arg\max_{\hat{y} \in Y} \{\sum_{\alpha \in A} \mu_{w,\alpha}(x, \widehat{y_\alpha}) + \sigma_{\mathrm{v}}(\mathrm{x}) \sum_{i=1}^{n} \gamma_i(\hat{y}_i) \}$$

- And define the loss-perturbed randomized predictor:

$$y^*_{w,\gamma,\boldsymbol{v}}(\epsilon) \in arg\, \max_{\hat{y} \in Y} \{\sum_{\alpha \in A} \mu_{w,\alpha}(x, \widehat{y_\alpha}) + \sigma_v(x) \sum_{i=1}^{n} \gamma_i(\hat{y}_i) + \epsilon \ell(y, \hat{y})\}$$

Integrating noise variance learning in direct loss minimization theorem

- The expected loss derivatives are:

$$\nabla_w E_\gamma\left[\ell\left(y, y^*_{w,\gamma,v}\right)\right] = \lim_{\epsilon\to 0}\frac{1}{\epsilon} E_\gamma\left[\sum_{\alpha\in A} \nabla_w \mu_{w,\alpha}\left(x, y^*_\alpha(\epsilon)\right) - \nabla_w \mu_{w,\alpha}(x, y^*_\alpha)\right]$$

$$\nabla_v E_\gamma\left[\ell\left(y, y^*_{w,\gamma,v}\right)\right] = \lim_{\epsilon\to 0}\frac{1}{\epsilon} E_\gamma\left[\textstyle\sum_{i=1}^{n} \nabla_v \sigma_v(x)\left(\gamma_i\left(y^*_i(\epsilon)\right) - \gamma_i\left(y^*_i\right)\right)\right]$$

Noise perturbation guarantees unique maximizers

- <u>Theorem</u>: Let $\gamma_{\mathrm{i}}(y_{\mathrm{i}})$ $be\ i.i.d\ random\ variables\ with$ a smooth probability density function. Then the set of maximal arguments of

$$y^*_{w,\gamma,v} = \arg\max_{\hat{y}\in Y} \{\sum_{\alpha\in A} \mu_{w,\alpha}(x, y_\alpha) + \sigma_{\mathrm{v}}(\mathrm{x}) \sum_{i=1}^{n} \gamma_i(y_i)\}$$

consists of a single structure with probability one for any $\gamma(y)$.

Experiments

- We validate the advantage of our approach in two popular structured prediction problems: bipartite matching and k-nearest neighbors.
- We compare to:
  - Direct loss minimization ($\overline{var} = 0$)
  - Lorberbom et al., 2018 ($\overline{var} = 1$)
  - State-of-the-art bipartite matching [Mena et al., 2018].
  - State-of-the-art neural sorting [Grover et al.,2019, Xie and Ermon, 2019].