Dropout Training as Adaptive Regularization

Stefan Wager, Sida Wang, and Percy Liang
Stanford University

Advances in Neural Information Processing Systems
December 6, 2013

Dropout acts as a *label-independent regularizer.*

- Dropout training solves:

$$\hat{\beta}_{DROPOUT} = \text{argmin}_\beta \left\{ \sum_{i=1}^{n} \mathbb{E}\left[\ell\left(\beta; \tilde{x}^{(i)}, y^{(i)} \right) \right] \right\},$$

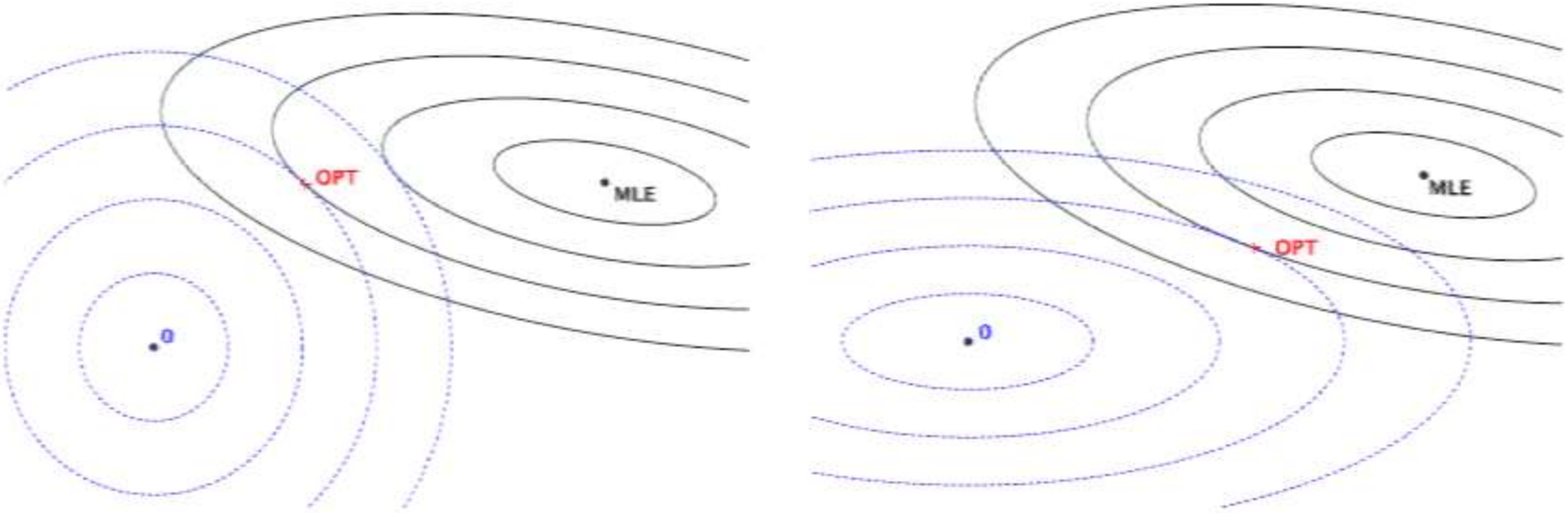
$$\text{where } \tilde{x}_j^{(i)} = \begin{cases} 0 & \text{with prob. } \delta \\ x_j^{(i)}/(1-\delta) & \text{with prob. } 1-\delta \end{cases}$$

- For generalized linear models (GLMs), this is equivalent to using a label-independent adaptive regularizer

$$\hat{\beta}_{DROPOUT} = \text{argmin}_\beta \left\{ \sum_{i=1}^{n} \ell\left(\beta; x^{(i)}, y^{(i)} \right) + R(\beta; x_i) \right\},$$

where $R(\cdot)$ is the dropout regularizer.

The shape of the dropout regularizer



(a) L_2 regularization (b) Dropout regularization

- For logistic regression, dropout favors confident predictions.
- Dropout is related to adaptive online learning schemes such as AdaGrad.

Semi-supervised dropout

- The dropout regularizer $R(\cdot)$ depends on features x but not on labels y, so we can use unlabeled data to improve $R(\cdot)$.
- Semi-supervised dropout improves on previous state-of-the-art results on the IMDB sentiment classification dataset of Maas et al (2011).