Dropout: Explicit Forms and Capacity Control

Raman Arora, Peter Bartlett, Poorya Mianjy, Nati Srebro

ICML

Introduction

- Algorithmic regularization techniques endow deep learning systems with reach inductive bias that helps them generalize.

Introduction

- Algorithmic regularization techniques endow deep learning systems with reach inductive bias that helps them generalize.
- We focus on dropout [Hinton et al., 2012], which is one of the most popular regularization techniques in today's practice of deep learning.

Introduction

- Algorithmic regularization techniques endow deep learning systems with reach inductive bias that helps them generalize.
- We focus on dropout [Hinton et al., 2012], which is one of the most popular regularization techniques in today's practice of deep learning.
- In particular, for matrix sensing and two-layer ReLU networks,
  - We analyze the *explicit forms* of the regularizer induced by dropout,
  - We demonstrate the *capacity control* due to dropout, by providing precise generalization error bounds.

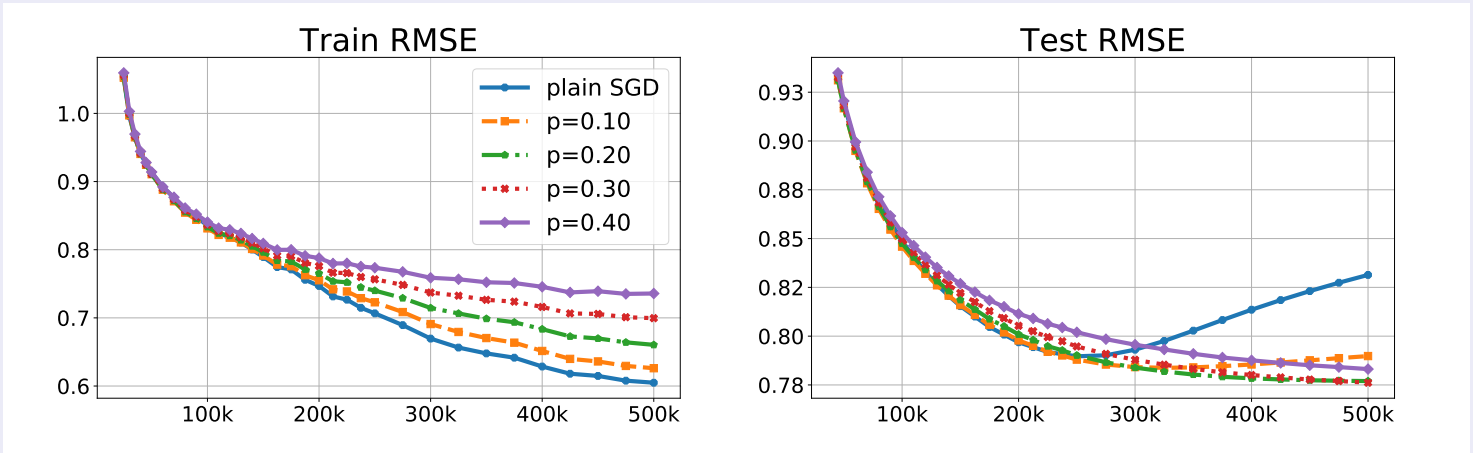
Dropout in Matrix Sensing

- *Goal:* recover $\mathrm{W}_* \in \mathbb{R}^{d_2 \times d_0}$ from n linear measurements $y_i = \langle \mathrm{X}_i, \mathrm{W}_* \rangle$
- Minimize squared loss in the factored form ($\mathrm{W}_2 \in \mathbb{R}^{d_2 \times d_1}$, $\mathrm{W}_1 \in \mathbb{R}^{d_1 \times d_0}$):

$$\min_{\mathrm{W}_2, \mathrm{W}_1} \hat{L}(\mathrm{W}_2, \mathrm{W}_1) := \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \mathrm{W}_2 \mathrm{W}_1, \mathrm{X}_i \rangle)^2$$

MovieLens-10M Dataset

- Collaborative filtering: $10M$ ratings for $11K$ movies by $72K$ users



Dropout in Matrix Sensing

- *Goal:* recover $\mathrm{W}_* \in \mathbb{R}^{d_2 \times d_0}$ from n linear measurements $y_i = \langle \mathrm{X}_i, \mathrm{W}_* \rangle$
- Minimize squared loss in the factored form ($\mathrm{W}_2 \in \mathbb{R}^{d_2 \times d_1}$, $\mathrm{W}_1 \in \mathbb{R}^{d_1 \times d_0}$):

$$\min_{\mathrm{W}_2, \mathrm{W}_1} \hat{L}(\mathrm{W}_2, \mathrm{W}_1) := \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \mathrm{W}_2 \mathrm{W}_1, \mathrm{X}_i \rangle)^2$$

MovieLens-10M Dataset

- Without *explicit* regularization, SGD suffers from gross overfitting
- Dropout consistently outperforms even with an early stopping oracle

	plain SGD		dropout			
width	last iterate	best iterate	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$
$m = 30$	0.8041	0.7938	0.7805	0.785	0.7991	0.8186
$m = 70$	0.8315	0.7897	0.7899	0.7771	0.7763	0.7833
$m = 110$	0.8431	0.7873	0.7988	0.7813	0.7742	0.7743
$m = 150$	0.8472	0.7858	0.8042	0.7852	0.7756	0.7722
$m = 190$	0.8473	0.7844	0.8069	0.7879	0.7772	0.772

Dropout in Matrix Sensing

- *Goal:* recover $\mathrm{W}_* \in \mathbb{R}^{d_2 \times d_0}$ from n linear measurements $y_i = \langle \mathrm{X}_i, \mathrm{W}_* \rangle$
- Minimize squared loss in the factored form ($\mathrm{W}_2 \in \mathbb{R}^{d_2 \times d_1}$, $\mathrm{W}_1 \in \mathbb{R}^{d_1 \times d_0}$):

$$\min_{\mathrm{W}_2, \mathrm{W}_1} \hat{L}(\mathrm{W}_2, \mathrm{W}_1) := \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \mathrm{W}_2 \mathrm{W}_1, \mathrm{X}_i \rangle)^2$$

$$\text{dropout objective} = \hat{L}(\mathrm{W}_2, \mathrm{W}_1) + \text{dropout regularizer}$$

Theorem (Explicit Form and Capacity Control)

X_i indicator matrix $\sim \mathbb{P}(row = i, col = j) = \mathbb{P}(row = i)\mathbb{P}(col = j) = p(i)q(j)$.

$$\textit{dropout regularizer} \propto \| \operatorname{diag}(\sqrt{\hat{p}})\, W_2\, W_1 \operatorname{diag}(\sqrt{\hat{q}}) \|_*^2$$

With probability $1 - \delta$ over the random draw of a sample of size n, the dropout rule output with dropout regularizer $\leq \alpha / d_1$ has generalization gap bounded as:

$$\text{generalization gap} \lesssim \sqrt{\frac{\alpha d_2 \log(d_2) + \log(1/\delta)}{n}}$$

2-layer ReLU Nets

- 2-layer ReLU nets, single ouput $d_2 = 1$, computing $f(\mathrm{x}; \mathrm{W}_2, \mathrm{W}_1) = \mathrm{W}_2\sigma(\mathrm{W}_1\mathrm{x})$.

2-layer ReLU Nets

- 2-layer ReLU nets, single ouput $d_2 = 1$, computing $f(\mathrm{x}; \mathrm{W}_2, \mathrm{W}_1) = \mathrm{W}_2\sigma(\mathrm{W}_1\mathrm{x})$.
- Class of networks with bounded dropout regularizer

$$\mathcal{H}_r := \{f(\cdot; \mathrm{w}),\ \text{dropout regularizer} \leq r\}$$

2-layer ReLU Nets

- 2-layer ReLU nets, single ouput $d_2 = 1$, computing $f(\mathrm{x}; \mathrm{W}_2, \mathrm{W}_1) = \mathrm{W}_2\sigma(\mathrm{W}_1\mathrm{x})$.
- Class of networks with bounded dropout regularizer

$$\mathcal{H}_r := \{f(\cdot; \mathrm{w}),\ \text{dropout regularizer} \leq r\}$$

- β-retentiveness: for any non-zero vector v, it holds that $\mathbb{E}\sigma(\mathrm{v}^\top \mathrm{x})^2 \geq \beta \mathbb{E}(\mathrm{v}^\top \mathrm{x})^2$

2-layer ReLU Nets

- 2-layer ReLU nets, single ouput $d_2 = 1$, computing $f(\mathrm{x}; \mathrm{W}_2, \mathrm{W}_1) = \mathrm{W}_2\sigma(\mathrm{W}_1\mathrm{x})$.
- Class of networks with bounded dropout regularizer

$$\mathcal{H}_r := \{f(\cdot; \mathrm{w}),\ \text{dropout regularizer} \leq r\}$$

- β-retentiveness: for any non-zero vector v, it holds that $\mathbb{E}\sigma(\mathrm{v}^\top \mathrm{x})^2 \geq \beta \mathbb{E}(\mathrm{v}^\top \mathrm{x})^2$

Theorem (Rademacher complexity - Upperbound)

For any sample S of size n, $\mathfrak{R}_S(\mathcal{H}_r) \leq \frac{2\sqrt{d_1 r}\|X\|_{C^\dagger}}{n\sqrt{\beta}}$.

2-layer ReLU Nets

- 2-layer ReLU nets, single ouput $d_2 = 1$, computing $f(\mathrm{x}; \mathrm{W}_2, \mathrm{W}_1) = \mathrm{W}_2\sigma(\mathrm{W}_1\mathrm{x})$.
- Class of networks with bounded dropout regularizer

$$\mathcal{H}_r := \{f(\cdot; \mathrm{w}),\ \text{dropout regularizer} \leq r\}$$

- β-retentiveness: for any non-zero vector v, it holds that $\mathbb{E}\sigma(\mathrm{v}^\top \mathrm{x})^2 \geq \beta \mathbb{E}(\mathrm{v}^\top \mathrm{x})^2$

Theorem (Rademacher complexity - Upperbound)

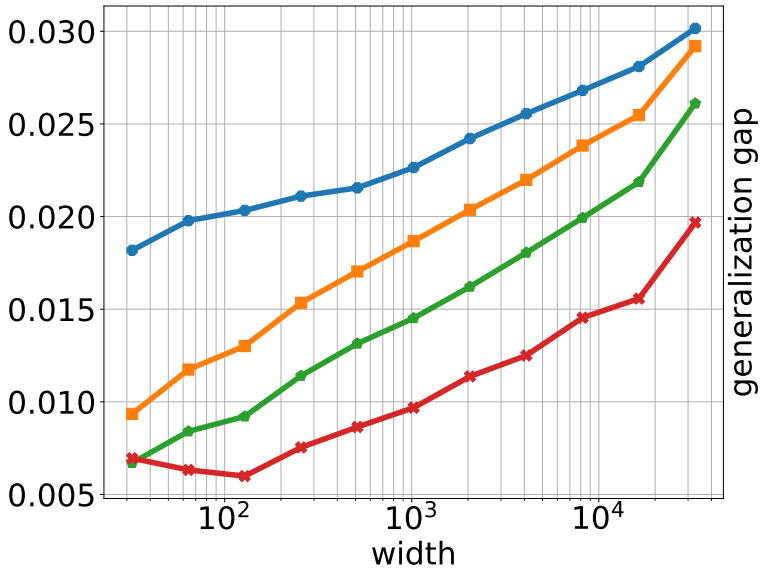
For any sample S of size n, $\mathfrak{R}_S(\mathcal{H}_r) \leq \frac{2\sqrt{d_1 r}\|X\|_{C^\dagger}}{n\sqrt{\beta}}$.

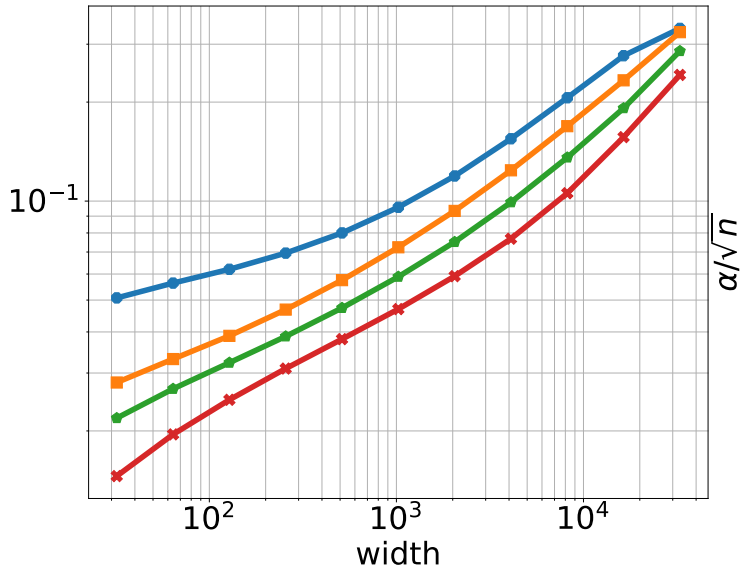
Theorem (Rademacher complexity - Lowerbound)

There is a constant c such that for any $r > 0$, $\mathfrak{R}_S(\mathcal{H}_r) \geq \frac{c\sqrt{d_1 r}\|X\|_{C^\dagger}}{n}$.

Empirical Results

- MNIST dataset of handwritten digits, extract two classes $\{4, 7\}$
- The trained networks achieve %100 training accuracy





Thanks for your attention!