Dropout Regularization Versus ℓ_2-Penalization in the Linear Model

Gabriel Clara Sophie Langer Johannes Schmidt-Hieber

Department of Applied Mathematics, Universiteit Twente

November 9, 2023

UNIVERSITY
OF TWENTE.

Joint work with Sophie Langer and Johannes Schmidt-Hieber

G.C., Sophie Langer, and Johannes Schmidt-Hieber. "Dropout Regularization Versus ℓ_2-Penalization in the Linear Model." *arXiv preprint: 2306.10529* (2023).

Dropout in Neural Networks

- Neural network with activation σ

$$f(x) = T_{W^{(L)},v^{(L)}} \circ \cdots \circ T_{W^{(1)},v^{(1)}}(x)$$

where $T_{W^{(\ell)},v^{(\ell)}} : z \mapsto \sigma\Big(W^{(\ell)}z + v^{(\ell)}\Big)$.

Dropout in Neural Networks

- Neural network with activation σ

$$f(x) = T_{W^{(L)},v^{(L)}} \circ \cdots \circ T_{W^{(1)},v^{(1)}}(x)$$

where $T_{W^{(\ell)},v^{(\ell)}} : z \mapsto \sigma\Big(W^{(\ell)}z + v^{(\ell)}\Big)$.

- During each iteration of training, dropout replaces each $T_{W^{(\ell)},v^{(\ell)}}$ with a sample from

$$z \mapsto \sigma\Big(W^{(\ell)}D^{(\ell)}z + v^{(\ell)}\Big)$$

where $D_{ii}^{(\ell)} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$.

Dropout in Neural Networks

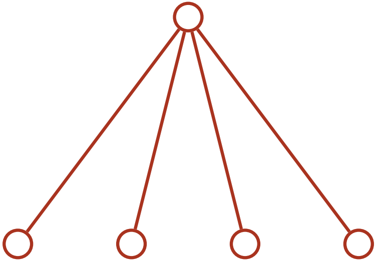


Figure: Regular neuron (left) and one sample of a neuron with dropout (right).

Why Study the Linear Model?

Canonical piece of wisdom: adding dropout noise to linear regression performs ridge regression/ℓ_2-penalization/Thikhonov regularization!

Why Study the Linear Model?

Proposition (Srivastava et al. Section 9)

Dropout matrix $D_{ii} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$*; linear model* $Y = X\beta_\star + \varepsilon$ *with standard normal noise independent of* D*, then*

$$\arg\min_{\beta} \mathbb{E}\Big[\|Y - XD\beta\|_2^2 \mid Y\Big] = \Big(pX^{\mathrm{t}}X + (1-p)\mathrm{Diag}\big(X^{\mathrm{t}}X\big)\Big)^{-1} X^{\mathrm{t}}Y$$

Why Study the Linear Model?

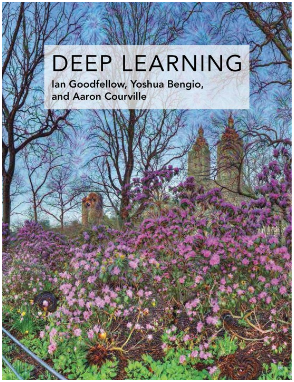
Proposition (Srivastava et al. Section 9)

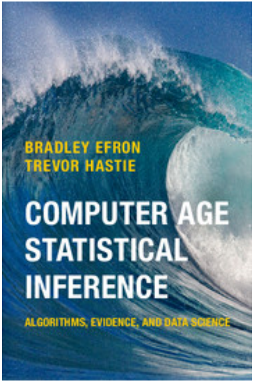
Dropout matrix $D_{ii} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$*; linear model* $Y = X\beta_\star + \varepsilon$ *with standard normal noise independent of* D*, then*

$$\arg\min_{\beta} \mathbb{E}\Big[\|Y - XD\beta\|_2^2 \mid Y\Big] = \Big(pX^{\mathrm{t}}X + (1-p)\mathrm{Diag}\big(X^{\mathrm{t}}X\big)\Big)^{-1} X^{\mathrm{t}}Y =: \tilde{\beta}$$

Why Study the Linear Model?

Canonical piece of wisdom: adding dropout noise to linear regression performs ridge regression/ℓ_2-penalization/Thikhonov regularization!





Some Definitions

- Important matrices:

$$\mathbb{X} := X^{\mathrm{t}}X$$
$$\overline{\mathbb{X}} := \mathbb{X} - \mathrm{Diag}(\mathbb{X})$$
$$\mathbb{X}_p := p\mathbb{X} + (1-p)\mathrm{Diag}(\mathbb{X})$$

Some Definitions

- Important matrices:

$$\mathbb{X} := X^{\mathrm{t}}X$$
$$\overline{\mathbb{X}} := \mathbb{X} - \mathrm{Diag}(\mathbb{X})$$
$$\mathbb{X}_p := p\mathbb{X} + (1-p)\mathrm{Diag}(\mathbb{X})$$

($\mathbb{X}_p$ invertible if $\min_i \mathbb{X}_{ii} > 0$)

Some Definitions

- Important matrices:

$$\mathbb{X} := X^{\mathrm{t}} X$$
$$\overline{\mathbb{X}} := \mathbb{X} - \mathrm{Diag}(\mathbb{X})$$
$$\mathbb{X}_p := p\mathbb{X} + (1-p)\mathrm{Diag}(\mathbb{X})$$

- *Averaged dropout estimator*: $\tilde{\beta} = \mathbb{X}_p^{-1} X^{\mathrm{t}} Y$ (minimizer from proposition)

Incorporating Dropout with Gradient Descent

Standard Gradient Descent:

$$\beta_{k+1} = \beta_k - \frac{\alpha}{2} \nabla_{\beta_k} \left\| Y - X\beta_k \right\|_2^2$$

Incorporating Dropout with Gradient Descent

Standard Gradient Descent:

$$\beta_{k+1} = \beta_k - \frac{\alpha}{2}\nabla_{\beta_k}\left\|Y - X\beta_k\right\|_2^2$$

On-Line Dropout:

$$\tilde{\beta}_{k+1} = \tilde{\beta}_k - \frac{\alpha}{2}\nabla_{\tilde{\beta}_k}\left\|Y - XD_{k+1}\tilde{\beta}_k\right\|_2^2$$

A new $i.i.d.$ dropout matrix is sampled every iteration!

Incorporating Dropout with Gradient Descent

On-Line Dropout:

$$\tilde{\beta}_{k+1} = \tilde{\beta}_k - \frac{\alpha}{2} \nabla_{\tilde{\beta}_k} \left\| Y - X D_{k+1} \tilde{\beta}_k \right\|_2^2$$

A new $i.i.d.$ dropout matrix is sampled every iteration!

Questions:

- Convergence towards $\tilde{\beta}$?
- Statistical optimality?

Convergence of Expectation

Proposition

If $\alpha p \|\mathbb{X}\| < 1$ *and* $\min_i \mathbb{X}_{ii} > 0$*, then*

$$\left\|\mathbb{E}[\tilde{\beta}_k - \tilde{\beta}]\right\|_2 \leq \left\|I - \alpha p \mathbb{X}_p\right\|^k \cdot \left\|\mathbb{E}[\tilde{\beta}_0 - \tilde{\beta}]\right\|_2$$

Second Moment Dynamics

Lemma

Up to exponentially decaying remainder ρ_k*, second moment of* $\tilde{\beta}_k - \tilde{\beta}$ *evolves as affine dynamical system*

$$\mathbb{E}\Big[(\tilde{\beta}_k - \tilde{\beta})(\tilde{\beta}_k - \tilde{\beta})^{\mathsf{t}}\Big] = S\Big(\mathbb{E}\Big[(\tilde{\beta}_{k-1} - \tilde{\beta})(\tilde{\beta}_{k-1} - \tilde{\beta})^{\mathsf{t}}\Big]\Big) + \rho_{k-1}$$

pushed forward by affine operator S *on matrices.*

Second Moment Dynamics

Lemma

Up to exponentially decaying remainder ρ_k, second moment of $\tilde{\beta}_k - \tilde{\beta}$ evolves as affine dynamical system

$$\mathbb{E}\Big[(\tilde{\beta}_k - \tilde{\beta})(\tilde{\beta}_k - \tilde{\beta})^{\mathrm{t}}\Big] = S\bigg(\mathbb{E}\Big[(\tilde{\beta}_{k-1} - \tilde{\beta})(\tilde{\beta}_{k-1} - \tilde{\beta})^{\mathrm{t}}\Big]\bigg) + \rho_{k-1}$$

pushed forward by affine operator S on matrices.

Intuition:

- Interaction between GD dynamics and on-line dropout encapsulated in S
- This structure remains hidden when considering averaged estimator $\tilde{\beta}$

Second Moment Dynamics

Lemma

Up to exponentially decaying remainder ρ_k, second moment of $\tilde{\beta}_k - \tilde{\beta}$ evolves as affine dynamical system

$$\mathbb{E}\Big[(\tilde{\beta}_k - \tilde{\beta})(\tilde{\beta}_k - \tilde{\beta})^{\mathrm{t}}\Big] = S\Big(\mathbb{E}\Big[(\tilde{\beta}_{k-1} - \tilde{\beta})(\tilde{\beta}_{k-1} - \tilde{\beta})^{\mathrm{t}}\Big]\Big) + \rho_{k-1}$$

pushed forward by affine operator S on matrices.

Exact Definition:

$$\begin{aligned} S(A) = {} & (I - \alpha p \mathbb{X}_p) A (I - \alpha p \mathbb{X}_p) + \alpha^2 p(1-p) \mathrm{Diag}(\mathbb{X}_p A \mathbb{X}_p) \\ & + \alpha^2 p^2 (1-p)^2 \overline{\mathbb{X}} \odot \left(A + \mathbb{E}[\tilde{\beta}\tilde{\beta}^{\mathrm{t}}]\right) \odot \overline{\mathbb{X}} \\ & + \alpha^2 p^2 (1-p) \Big(\overline{\mathbb{X}} \mathrm{Diag}\Big(A + \mathbb{E}[\tilde{\beta}\tilde{\beta}^{\mathrm{t}}]\Big)\overline{\mathbb{X}}\Big)_p \\ & + \alpha^2 p^2 (1-p) \Big(\overline{\mathbb{X}} \mathrm{Diag}(\mathbb{X}_p A) + \mathrm{Diag}(\mathbb{X}_p A)\overline{\mathbb{X}}\Big) \end{aligned}$$

Convergence of Variance

Theorem

For sufficiently small $\alpha := \alpha(\mathbb{X}, p)$, $S_0 := S(0)$, *and* $S_{\text{lin}} := S - S_0$

$$\operatorname{Cov}\big(\tilde{\beta}_k\big) = \operatorname{Cov}\big(\tilde{\beta}\big) + \big(\mathrm{id} - S_{\text{lin}}\big)^{-1} S_0 + O\Big(k \|I - \alpha p \mathbb{X}_p\|^{k-1}\Big)$$

Convergence of Variance

Theorem

For sufficiently small $\alpha := \alpha(\mathbb{X}, p)$, $S_0 := S(0)$, *and* $S_{\mathrm{lin}} := S - S_0$

$$\mathrm{Cov}(\tilde{\beta}_k) = \mathrm{Cov}(\tilde{\beta}) + \left(\mathrm{id} - S_{\mathrm{lin}}\right)^{-1} S_0 + O\Big(k \|I - \alpha p \mathbb{X}_p\|^{k-1}\Big)$$

Notes:

- Limit characterized by intercept S_0 and linear part S_{lin} of S

Convergence of Variance

Theorem

For sufficiently small $\alpha := \alpha(\mathbb{X}, p)$, $S_0 := S(0)$, *and* $S_{\text{lin}} := S - S_0$

$$\mathrm{Cov}(\tilde{\beta}_k) = \mathrm{Cov}(\tilde{\beta}) + \left(\mathrm{id} - S_{\text{lin}}\right)^{-1} S_0 + O\Big(k \|I - \alpha p \mathbb{X}_p\|^{k-1}\Big)$$

Corollary:

- Unfortunately, $\left(\mathrm{id} - S_{\text{lin}}\right)^{-1} S_0 \neq 0$ in general, so

$$\mathrm{Tr}\Big(\mathbb{E}\Big[(\tilde{\beta}_k - \tilde{\beta})(\tilde{\beta}_k - \tilde{\beta})^{\mathsf{t}}\Big]\Big) = \mathbb{E}\Big[\|\tilde{\beta}_k - \tilde{\beta}\|_2^2\Big] > 0.$$

For more details:
G.C., Sophie Langer, and Johannes Schmidt-Hieber. "Dropout Regularization Versus ℓ_2-Penalization in the Linear Model." *arXiv preprint: 2306.10529* (2023).

For more details:
G.C., Sophie Langer, and Johannes Schmidt-Hieber. “Dropout Regularization Versus ℓ_2-Penalization in the Linear Model.” *arXiv preprint: 2306.10529* (2023).

Thanks for your attention!