Dropout Regularization Versus ℓ_2-Penalization in the Linear Model

Gabriel Clara Sophie Langer Johannes Schmidt-Hieber

Department of Applied Mathematics, Universiteit Twente

March 15, 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

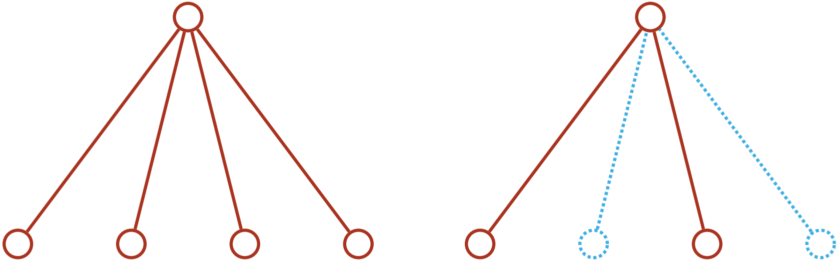

Figure: Regular neuron (left) and one sample of a neuron with dropout (right).

Why Study the Linear Model?

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

Why Study the Linear Model?

Proposition (Srivastava et al.)

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.[a])

[a] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R, Salakhutdinov. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting.* JMLR. 2014.

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.)

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?

Proposition (Srivastava et al.)

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

$$\operatorname*{arg\,min}_{\beta} \mathbb{E}\Big[\|Y - XD\beta\|_2^2 \mid Y\Big] =: \tilde{\beta}$$

Intuition:

- Re-scaled minimizer of the averaged loss performs weighted ridge regression:

$$p\tilde{\beta} = \operatorname*{arg\,min}_{\beta} \left(\|Y - X\beta\|_2^2 + \left(\frac{1}{p} - 1\right) \cdot \left\|\sqrt{\mathrm{Diag}(X^{\mathrm{t}}X)}\beta\right\|_2^2\right)$$

- Small $p \implies$ strong regularization

Why Study the Linear Model?

Proposition (Srivastava et al.)

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*

$$\arg\min_{\beta} \mathbb{E}\Big[\|Y - XD\beta\|_2^2 \mid Y\Big] =: \tilde{\beta}$$

Problems:

- No explicit gradient descent
- No access to variance

Why Study the Linear Model?

Proposition (Srivastava et al.)

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*

$$\arg\min_\beta \mathbb{E}\Big[\|Y - XD\beta\|_2^2 \mid Y\Big] =: \tilde{\beta}$$

Problems:

- No explicit gradient descent
- No access to variance
- Conditional expectation $\mathbb{E}[\,\cdot \mid Y]$ represents loss of information $\implies \tilde{\beta}$ may not fully capture gradient descent dynamics

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*

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

Convergence of Expectation

Proposition

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

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

Intuition:

- Exponential decay, as in regular gradient descent
- Expected learning rate αp

Convergence of Expectation

Proposition

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

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

Idea of proof:

- Rewrite

$$\tilde{\beta}_k - \tilde{\beta} = (I - \alpha D_k \mathbb{X} D_k)(\tilde{\beta}_{k-1} - \tilde{\beta}) + \alpha D_k \overline{\mathbb{X}}(pI - D_k)\tilde{\beta}$$

Convergence of Expectation

Proposition

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

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

Idea of proof:

- Rewrite

$$\tilde{\beta}_k - \tilde{\beta} = \big(I - \alpha D_k \mathbb{X} D_k\big)\big(\tilde{\beta}_{k-1} - \tilde{\beta}\big) + \alpha D_k \overline{\mathbb{X}}\big(pI - D_k\big)\tilde{\beta}$$

- Compute

$$\mathbb{E}\big[D_k \mathbb{X} D_k\big] = p\mathbb{X}_p$$
$$\mathbb{E}\big[D_k \overline{\mathbb{X}}(pI - D_k)\big] = 0$$

Convergence of Expectation

Proposition

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

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

Idea of proof:

- Rewrite

$$\tilde{\beta}_k - \tilde{\beta} = (I - \alpha D_k \mathbb{X} D_k)(\tilde{\beta}_{k-1} - \tilde{\beta}) + \alpha D_k \overline{\mathbb{X}}(pI - D_k)\tilde{\beta}$$

- Compute

$$\mathbb{E}\big[D_k \mathbb{X} D_k\big] = p\mathbb{X}_p$$
$$\mathbb{E}\big[D_k \overline{\mathbb{X}}(pI - D_k)\big] = 0$$

- Now $\mathbb{E}[\tilde{\beta}_k - \tilde{\beta}] = (I - \alpha p \mathbb{X}_p)\mathbb{E}[\tilde{\beta}_{k-1} - \tilde{\beta}]$; finish with induction!

Second Moment Dynamics I

Theorem (Informal Statement)

Affine estimator $\tilde{\beta}_{\mathrm{aff}} := BY + a$ *(with* B *and* a *independent of* Y*) and linear estimator* $\tilde{\beta}_A := AX^{\mathrm{t}}Y$ *(with* A *deterministic), then*

$$\mathbb{E}\big[\tilde{\beta}_{\mathrm{aff}}\big] \approx \mathbb{E}\big[\tilde{\beta}_A\big] \implies \mathrm{Cov}\big(\tilde{\beta}_{\mathrm{aff}} - \tilde{\beta}_A, \tilde{\beta}_A\big) \approx 0$$

Second Moment Dynamics I

Theorem (Informal Statement)

Affine estimator $\tilde{\beta}_{\mathrm{aff}} := BY + a$ *(with* B *and* a *independent of* Y*) and linear estimator* $\tilde{\beta}_A := AX^{\mathrm{t}}Y$ *(with* A *deterministic), then*

$$\mathbb{E}\big[\tilde{\beta}_{\mathrm{aff}}\big] \approx \mathbb{E}\big[\tilde{\beta}_A\big] \implies \mathrm{Cov}\big(\tilde{\beta}_{\mathrm{aff}} - \tilde{\beta}_A, \tilde{\beta}_A\big) \approx 0$$

Intuition:

- If $\tilde{\beta}_{\mathrm{aff}}$ is (nearly) unbiased for $\tilde{\beta}_A$, then

$$\tilde{\beta}_{\mathrm{aff}} \approx \tilde{\beta}_A + \text{centered orthogonal noise}$$

Second Moment Dynamics I

Theorem (Informal Statement)

Affine estimator $\tilde{\beta}_{\text{aff}} := BY + a$ *(with* B *and* a *independent of* Y*) and linear estimator* $\tilde{\beta}_A := AX^{\text{t}}Y$ *(with* A *deterministic), then*

$$\mathbb{E}[\tilde{\beta}_{\text{aff}}] \approx \mathbb{E}[\tilde{\beta}_A] \implies \operatorname{Cov}(\tilde{\beta}_{\text{aff}} - \tilde{\beta}_A, \tilde{\beta}_A) \approx 0$$

Intuition:

- If $\tilde{\beta}_{\text{aff}}$ is (nearly) unbiased for $\tilde{\beta}_A$, then

$$\tilde{\beta}_{\text{aff}} \approx \tilde{\beta}_A + \text{centered orthogonal noise}$$

- Gauss-Markov like corollary; if $B_k Y + a_k$ asymptotically unbiased for $\tilde{\beta}_A$, then

$$\liminf_{k \to \infty} \operatorname{Cov}(B_k Y + a_k) \geq \operatorname{Cov}(\tilde{\beta}_A)$$

Second Moment Dynamics I

Theorem (Informal Statement)

Affine estimator $\tilde{\beta}_{\mathrm{aff}} := BY + a$ (with B and a independent of Y) and linear estimator $\tilde{\beta}_A := AX^{\mathrm{t}}Y$ (with A deterministic), then

$$\mathbb{E}\big[\tilde{\beta}_{\mathrm{aff}}\big] \approx \mathbb{E}\big[\tilde{\beta}_A\big] \implies \mathrm{Cov}\big(\tilde{\beta}_{\mathrm{aff}} - \tilde{\beta}_A, \tilde{\beta}_A\big) \approx 0$$

Dropout-specific:

- Dropout iterates $\tilde{\beta}_k$ are affine estimators asymptotically unbiased for $\tilde{\beta}$
- $\mathrm{Cov}\big(\tilde{\beta}\big)$ represents fundamental lower bound

Second Moment Dynamics II

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.

Second Moment Dynamics II

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.

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 II

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:

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

Second Moment Dynamics II

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.*

Notes on Proof:

- S has complex expression due to dependence structure in

$$\tilde{\beta}_k - \tilde{\beta} = (I - \alpha D_k \mathbb{X} D_k)(\tilde{\beta}_{k-1} - \tilde{\beta}) + \alpha D_k \overline{\mathbb{X}}(pI - D_k)\tilde{\beta}$$

- Proof requires computing 4^{th} order moments of the form $\mathbb{E}\big[D_k A D_k B D_k C D_k\big]$

Second Moment Dynamics III

Theorem

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

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

Second Moment Dynamics III

Theorem

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

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

Notes:

- Limit characterized by intercept S_0 and linear part S_{lin} of S
- Small $\alpha \implies$ operator norm of S_{lin} less than 1

Second Moment Dynamics III

Theorem

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

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

Corollary I:

- $\mathrm{Cov}(\tilde{\beta}_k) = \mathrm{Cov}(\tilde{\beta}) + (\mathrm{id} - S_{\text{lin}})^{-1} S_0 + O\Big(k \| I - \alpha p \mathbb{X}_p \|^{k-1} \Big)$
- $(\mathrm{id} - S_{\text{lin}})^{-1} S_0$ is the variance of the “centered orthogonal noise” from earlier proposition

Second Moment Dynamics III

Theorem

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

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

Corollary I:

- $\mathrm{Cov}\left(\tilde{\beta}_k\right) = \mathrm{Cov}\left(\tilde{\beta}\right) + \left(\mathrm{id} - S_{\mathrm{lin}} \right)^{-1} S_0 + O\left(k \| I - \alpha p \mathbb{X}_p \|^{k-1} \right)$
- Unfortunately, $\left(\mathrm{id} - S_{\mathrm{lin}} \right)^{-1} S_0 \neq 0$ in general, so $\tilde{\beta}_k$ does not attain the optimal variance!

Second Moment Dynamics III

Theorem

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

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

Corollary I:

- $\mathrm{Cov}(\tilde{\beta}_k) = \mathrm{Cov}(\tilde{\beta}) + (\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0 + O\Big(k \| I - \alpha p \mathbb{X}_p \|^{k-1} \Big)$
- Unfortunately, $(\mathrm{id} - S_{\mathrm{lin}})^{-1} S_0 \neq 0$ in general, so $\tilde{\beta}_k$ does not attain the optimal variance!

Corollary II:

- In general, $\tilde{\beta}_k$ does not converge to $\tilde{\beta}$ in L_2 since

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

Conclusion

Our techniques/results show:

- Second-order analysis of gradient descent with dropout is already rather technical in the linear model.

Conclusion

Our techniques/results show:

- Second-order analysis of gradient descent with dropout is already rather technical in the linear model.
- Elementary — yet complicated — linear algebra is necessary at first to compute the basic objects, then a more abstract perspective can be applied.

Conclusion

Our techniques/results show:

- Second-order analysis of gradient descent with dropout is already rather technical in the linear model.
- Elementary — yet complicated — linear algebra is necessary at first to compute the basic objects, then a more abstract perspective can be applied.
- Second-order dynamics are only visible through direct study of on-line iterates.

Conclusion

Our techniques/results show:

- Second-order analysis of gradient descent with dropout is already rather technical in the linear model.
- Elementary — yet complicated — linear algebra is necessary at first to compute the basic objects, then a more abstract perspective can be applied.
- Second-order dynamics are only visible through direct study of on-line iterates.
- Often cited connection with ridge regression is more nuanced for the variance.

Extensions/Open Problems

- Neural networks?
- Connections with other forms of algorithmic regularization?
- Randomized design and iteration dependent learning rate?

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!