Dropout Regularization Versus ℓ_2-Penalization in the Linear Model

Gabriel Clara Sophie Langer Johannes Schmidt-Hieber

Department of Applied Mathematics, Universiteit Twente

August 24, 2023

UNIVERSITY OF TWENTE.

10th International Congress on Industrial and Applied Mathematics

ICIAM 2023 TOKYO

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).

Motivation: Model Averaging

Why perform model averaging in neural networks?

Motivation: Model Averaging

Why perform model averaging in neural networks?

- Breaking co-adaptation between neurons
- Preventing over-fitting

Motivation: Model Averaging

Why perform model averaging in neural networks?

- Breaking co-adaptation between neurons
- Preventing over-fitting

Ideally, Bayesian model averaging:

- Take prior on possible connections in the network
- Use posterior to average

Motivation: Model Averaging

Why perform model averaging in neural networks?

- Breaking co-adaptation between neurons
- Preventing over-fitting

Ideally, Bayesian model averaging:

- Take prior on possible connections in the network
- Use posterior to average

Problem: Combinatorial Explosion!

Dropout in Neural Networks

Proposed Solution: Dropout![1]

[1]Srivastava N. et al. "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research* (2014).

Dropout in Neural Networks

Proposed Solution: Dropout![1]

- Randomly exclude connections from training at every step of the gradient descent
- Re-scale trained weights appropriately

$\Longrightarrow$ Approximates model averaging while being tractable

[1]Srivastava N. et al. "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research* (2014).

Dropout in Neural Networks

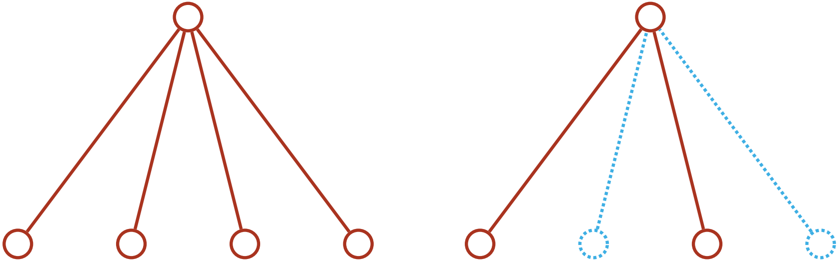


Figure: Regular neuron (left) and one sample of a neuron with dropout (right).

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?

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

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?

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

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*

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

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*

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

Intuition:

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

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

- Small $p \implies$ strong 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*

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

Problems:

- No explicit gradient descent
- No access to variance $\implies$ no statistical analysis

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*

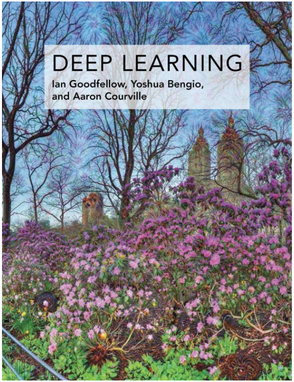
$$\arg\min_\beta \mathbb{E}\Big[\|Y - XD\beta\|_2^2 \mid Y\Big] =: \tilde{\beta}$$

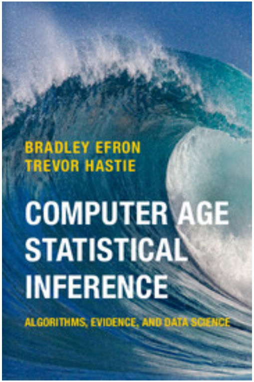
Problems:

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

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

- *Dropout matrix*: $D_{ii} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$

Some Definitions

- *Dropout matrix*: $D_{ii} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$
- Important matrices:

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

Some Definitions

- *Dropout matrix*: $D_{ii} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$
- Important matrices:

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

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

Some Definitions

- *Dropout matrix*: $D_{ii} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$
- Important matrices:

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

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

Some Definitions

- *Dropout matrix*: $D_{ii} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$
- Important matrices:

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

- *Averaged dropout estimator*: $\tilde{\beta} = \mathbb{X}_p^{-1} X^{\mathrm{t}} Y$
- Euclidean norm $\|\cdot\|_2$ on vectors; spectral norm $\|\cdot\|$ on matrices

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

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

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

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

- Compute

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

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

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

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

Second-Order Dynamics I

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-Order Dynamics I

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-Order Dynamics I

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-Order Dynamics I

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.

Notes on Proof:

- S has complex expression due to dependence structure in

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

- 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-Order Dynamics II

Theorem

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

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

Second-Order Dynamics II

Theorem

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

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

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-Order Dynamics II

Theorem

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

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

Corollary I:

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

Second-Order Dynamics II

Theorem

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

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

Corollary I:

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

Second-Order Dynamics II

Theorem

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

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

Corollary I:

- $\mathrm{Cov}\big(\tilde{\beta}_k\big) = \mathrm{Cov}\big(\tilde{\beta}\big) + \big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0 + O\Big(k \| I - \alpha p \mathbb{X}_p \|^{k-1} \Big)$
- Unfortunately, $\big(\mathrm{id} - S_{\mathrm{lin}}\big)^{-1} S_0 > 0$ 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.

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!