Dropout Regularization Versus ℓ_2-Penalization in the Linear Model

Gabriel Clara Sophie Langer Johannes Schmidt-Hieber

June 28, 2024

Department of Applied Mathematics, Universiteit 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 shifted activation $\sigma_v = \sigma(\,\cdot\, - v)$

$$f(x) = W^{(L)} \circ \sigma_{v^{(L)}} \circ \cdots \circ W^{(1)} \circ \sigma_{v^{(1)}} \circ W^{(0)}(x)$$

Dropout in Neural Networks

- Neural network with shifted activation $\sigma_v = \sigma(\,\cdot\, - v)$

$$f(x) = W^{(L)} \circ \sigma_{v^{(L)}} \circ \cdots \circ W^{(1)} \circ \sigma_{v^{(1)}} \circ W^{(0)}(x)$$

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

$$W^{(\ell)} D^{(\ell)}, \qquad D^{(\ell)}_{ii} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$$

Dropout in Neural Networks

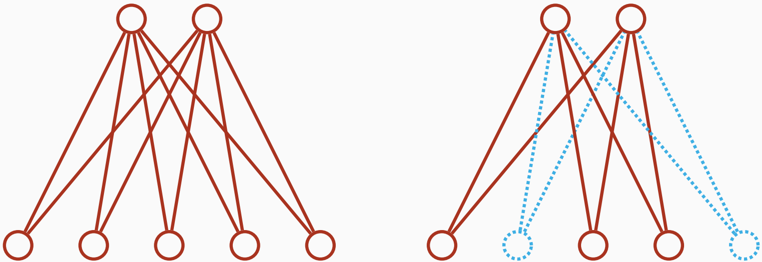

Figure 1: Regular neurons (left) and random sample of dropout neurons (right).

Why Study the Linear Model?

Canonical piece of wisdom: integrating over dropout noise in linear regression leads to ridge regression/ℓ_2-penalization!

Why Study the Linear Model?

Proposition (Srivastava et al.)

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

$$\arg\min_\beta \mathbb{E}_D\left[\|Y - XD\beta\|_2^2\right] = \left(pX^tX + (1-p)\mathrm{Diag}(X^tX)\right)^{-1} X^tY$$

Why Study the Linear Model?

Proposition (Srivastava et al.[1])

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

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

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

Why Study the Linear Model?

Proposition (Srivastava et al.)

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

$$\arg\min_\beta \mathbb{E}_D\left[\|Y - XD\beta\|_2^2\right] = \underbrace{\left(pX^tX + (1-p)\mathrm{Diag}\left(X^tX\right)\right)^{-1} X^tY}_{=:\widetilde{\beta}}$$

Why Study the Linear Model?

Proposition (Srivastava et al.)

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

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

Intuition:

- Re-scaled minimizer performs weighted ridge regression:

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

- 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* $Y = X\beta_\star + \varepsilon$*, then*

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

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* $Y = X\beta_\star + \varepsilon$*, then*

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

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 with extra noise

Some Definitions

- 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

- 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:

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

- *Marginalized dropout estimator*: $\tilde{\beta} = \mathbb{X}_p^{-1}X^{\mathsf{t}}Y$ (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:

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

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 $\tilde{\beta}$?
- Characterizing dynamics with noise?

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

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

- Compute

$$\mathbb{E}\left[D_k \mathbb{X} D_k\right] = p\mathbb{X}_p$$

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

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

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

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

Dropout in Neural Networks

Linear Regression as a Toy Model

Gradient Descent with Dropout

Second Moment Dynamics

Second Moment Dynamics I

Theorem (Informal Statement)

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

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

Second Moment Dynamics I

Theorem (Informal Statement)

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

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

Intuition:

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

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

Theorem (Informal Statement)

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

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

- If $B_kY + a_k$ asymptotically unbiased for $\tilde{\beta}_A$,

$$\liminf_{k\to\infty} \mathrm{Cov}\big(B_kY + a_k\big) \geq \mathrm{Cov}\big(\tilde{\beta}_A\big)$$

Second Moment Dynamics I

Theorem (Informal Statement)

Affine estimator $\tilde{\beta}_{\mathrm{aff}} := BY + a$ *(*B *and* a *independent of* Y*) and linear estimator* $\tilde{\beta}_A := AX^{\mathrm{t}}Y$ *(*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 $\tilde{\beta}$
- $\mathrm{Cov}(\tilde{\beta})$ represents fundamental lower bound

Second Moment Dynamics II

Lemma

Second moment of $\tilde{\beta}_k - \tilde{\beta}$ evolves as affine dynamical system

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

pushed forward by affine map S with decaying remainder ρ_k.

Second Moment Dynamics II

Lemma

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 map S with decaying remainder ρ_k.

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

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 map S with decaying remainder ρ_k.

Exact Definition:

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

Second Moment Dynamics II

Lemma

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 map S with decaying remainder ρ_k.

Notes on Proof:

- Complicated expression due to dependence structure in

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

Second Moment Dynamics II

Lemma

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 map S with decaying remainder ρ_k.

Notes on Proof:

- Complicated expression due to dependence structure in

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

- Requires computing 4th order moments $\mathbb{E}\big[D_k A D_k B D_k C D_k\big]$

Second Moment Dynamics III

Theorem

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

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

Second Moment Dynamics III

Theorem

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

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

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 Moment Dynamics III

Theorem

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

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

Corollary I:

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

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

Corollary I:

- $\mathrm{Cov}(\tilde{\beta}_k) = \mathrm{Cov}(\tilde{\beta}) + \left(\mathrm{id} - S_{\text{lin}} \right)^{-1} S_0 + O\left(k \| I - \alpha p \mathbb{X}_p \|^{k-1} \right)$
- $\left(\mathrm{id} - S_{\text{lin}} \right)^{-1} S_0$ is the variance of the "centered orthogonal noise" from earlier proposition
- Unfortunately, $\left(\mathrm{id} - S_{\text{lin}} \right)^{-1} S_0 \neq 0$ 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* $S_{\text{lin}} =: S - S_0$

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

Corollary II:

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

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

$$\to \mathrm{Tr}\left(\left(\mathrm{id} - S_{\text{lin}}\right)^{-1} S_0 \right).$$

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!

Sub-Optimality of Variance

Theorem

Suppose $\sup_{m \neq \ell} |\mathbb{X}_{\ell m}| \neq 0$ *for every* $\ell = 1, \ldots, d$*, then*

$$\lim_{k \to \infty} \mathrm{Cov}(\tilde{\beta}_k) - \mathrm{Cov}(\tilde{\beta}) \geq O\Big(\lambda_{\min}(\mathbb{X}) \min_{i \neq j : \mathbb{X}_{ij} \neq 0} \mathbb{X}_{ij}^2\Big) \cdot I_d$$

whenever the limit exists.

Sub-Optimality of Variance

Theorem

Suppose $\sup_{m \neq \ell} |\mathbb{X}_{\ell m}| \neq 0$ *for every* $\ell = 1, \ldots, d$*, then*

$$\lim_{k \to \infty} \mathrm{Cov}(\tilde{\beta}_k) - \mathrm{Cov}(\tilde{\beta}) \geq O\Big(\lambda_{\min}(\mathbb{X}) \min_{i \neq j : \mathbb{X}_{ij} \neq 0} \mathbb{X}_{ij}^2\Big) \cdot I_d$$

whenever the limit exists.

Notes:

- Non-trivial bound provided $\lambda_{\min}(\mathbb{X}) > 0$.

Sub-Optimality of Variance

Theorem

Suppose $\sup_{m \neq \ell} |\mathbb{X}_{\ell m}| \neq 0$ *for every* $\ell = 1, \ldots, d$*, then*

$$\lim_{k \to \infty} \mathrm{Cov}(\tilde{\beta}_k) - \mathrm{Cov}(\tilde{\beta}) \geq O\Big(\lambda_{\min}(\mathbb{X}) \min_{i \neq j : \mathbb{X}_{ij} \neq 0} \mathbb{X}_{ij}^2\Big) \cdot I_d$$

whenever the limit exists.

Notes:

- Non-trivial bound provided $\lambda_{\min}(\mathbb{X}) > 0$.
- Frobenius norm of right-hand side scales with dimension d.

Ruppert-Polyak Averaging

Theorem

Running average $\tilde{\beta}_k^{\mathrm{rp}} := \frac{1}{k}\sum_{\ell=1}^{k}\tilde{\beta}_\ell$; *for sufficiently small* α

$$\left\|\mathbb{E}\left[(\tilde{\beta}_k^{\mathrm{rp}} - \tilde{\beta})(\tilde{\beta}_k^{\mathrm{rp}} - \tilde{\beta})^{\mathsf{t}}\right]\right\| = O\left(k^{-1}\right)$$

Ruppert-Polyak Averaging

Theorem

Running average $\tilde{\beta}_k^{\mathrm{rp}} := \frac{1}{k}\sum_{\ell=1}^{k}\tilde{\beta}_\ell$*; for sufficiently small* α

$$\left\|\mathbb{E}\left[(\tilde{\beta}_k^{\mathrm{rp}} - \tilde{\beta})(\tilde{\beta}_k^{\mathrm{rp}} - \tilde{\beta})^{\mathrm{t}}\right]\right\| = O\left(k^{-1}\right)$$

Intuition:

- "Centered orthogonal noise" averaged away; at the price of slower convergence
- $\tilde{\beta}_k^{\mathrm{rp}}$ converges to $\tilde{\beta}$ in L_2