Coffee Talk #4

November 9, 2021

Özgür Taylan Turan

Learning Rate Annealing Can Provably Help Generalization Even For Convex Problems[1]

[1] P. Nakkiran (2020). "Learning Rate Annealing Can Provably Help Generalization, Even for Convex Problems". In: pp. 1–9. arXiv: 2005.07360

Why This Paper?

Simple, but not obvious observations...

Aim

Show that large initial learning rate can act as a regularizer!

- Non-convex setting [1]
- Convex setting [2]

[1] Y. Li, C. Wei, and T. Ma (2019). "Towards explaining the regularization effect of initial large learning rate in training neural networks". In: *Advances in Neural Information Processing Systems* 32, pp. 1–49. ISSN: 10495258. arXiv: 1907.04595

[2] P. Nakkiran (2020). "Learning Rate Annealing Can Provably Help Generalization, Even for Convex Problems". In: pp. 1–9. arXiv: 2005.07360

Non-Convex Setting [1]-A

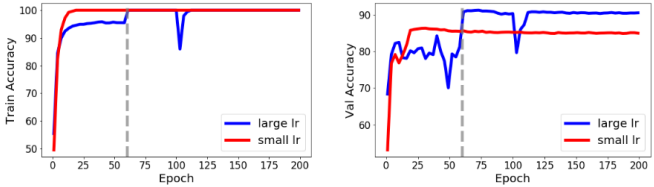


Figure 1: CIFAR-10 accuracy vs. epoch for WideResNet with weight decay, no data augmentation, and initial lr of 0.1 vs. 0.01. Gray represents the annealing time. **Left:** Train. **Right:** Validation.

[1] Y. Li, C. Wei, and T. Ma (2019). "Towards explaining the regularization effect of initial large learning rate in training neural networks". In: *Advances in Neural Information Processing Systems* 32, pp. 1–49. ISSN: 10495258. arXiv: 1907.04595

Non-Convex Setting [1]-B

Claim 1: Small learning rate $\rightarrow$ easy-to-generalize and hard-to-fit patterns
Claim 2: Large learning rate $\rightarrow$ hard-to-generalize and easy-to-fit patterns

Complex theoretical and a small empirical investigation.

[1] Y. Li, C. Wei, and T. Ma (2019). "Towards explaining the regularization effect of initial large learning rate in training neural networks". In: *Advances in Neural Information Processing Systems* 32, pp. 1–49. ISSN: 10495258. arXiv: 1907.04595

Non-Convex Setting [1]-C

Figure 4: Visualizations of CIFAR-10 images with patches added.

CIFAR-10: 20% No-patch, 16% Only-patch, 60% Image-with-patch

[1] Y. Li, C. Wei, and T. Ma (2019). "Towards explaining the regularization effect of initial large learning rate in training neural networks". In: *Advances in Neural Information Processing Systems* 32, pp. 1–49. ISSN: 10495258. arXiv: 1907.04595

Non-Convex Setting [1]-D

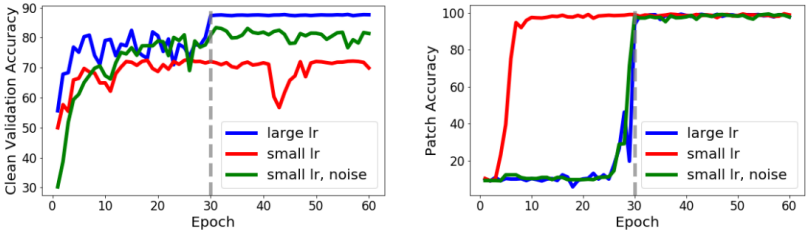


Figure 3: Accuracy vs. epoch on patch-augmented CIFAR-10. The gray line indicates annealing of activation noise and learning rate. **Left:** Clean validation set. **Right:** Images containing only the patch.

Small Gaussian-noise addition before activation layer have a regularizing effect on the small learning rate!

[1] Y. Li, C. Wei, and T. Ma (2019). "Towards explaining the regularization effect of initial large learning rate in training neural networks". In: *Advances in Neural Information Processing Systems* 32, pp. 1–49. ISSN: 10495258. arXiv: 1907.04595

Non-Convex Setting [1]-E

Claim 3: Convex problems have unique minimum, so this effect cannot be observed!

[1] Y. Li, C. Wei, and T. Ma (2019). "Towards explaining the regularization effect of initial large learning rate in training neural networks". In: *Advances in Neural Information Processing Systems* 32, pp. 1–49. ISSN: 10495258. arXiv: 1907.04595

Convex Setting [1]

Problem

- Assume a distribution $\mathcal{D}$ over $(x, y) \in \mathbb{R}^2 x \mathbb{R}$
- Given $x \in \{\mathbf{e}_1, \mathbf{e}_2\}$ uniformly at random; $y = \langle \beta^*, x \rangle$ for ground truth $\mathbb{R}^2$
- Learn a linear model $\hat{y} := \langle \beta, x \rangle$, where $\beta = (\beta_1, \beta_2)$
- Noting, population loss $\rightarrow L_{\mathcal{D}} := \mathbb{E}_D \left[(\langle \beta, x \rangle - y)^2 \right]$
- And, empirical loss $\rightarrow \hat{L}_n := \frac{1}{n} \sum_i (\langle \beta, x_i \rangle - y_i)^2$ for drawn n samples from the distribution $\mathcal{D}$

[1] P. Nakkiran (2020). "Learning Rate Annealing Can Provably Help Generalization, Even for Convex Problems". In: pp. 1–9. arXiv: 2005.07360

Convex Setting [1]

Assume you have $n = 3$ with $(x_i, y_i)_{i=1}^3 \to \{(\mathbf{e}_1, \beta_1^*), (\mathbf{e}_1, \beta_1^*), (\mathbf{e}_2, \beta_2^*)\}$

Empirical Loss

$$\hat{L}_n := \tfrac{2}{3}(\beta_1 - \beta_1^*)^2 + \tfrac{1}{3}(\beta_2 - \beta_2^*)^2$$

[1] P. Nakkiran (2020). "Learning Rate Annealing Can Provably Help Generalization, Even for Convex Problems". In: pp. 1–9. arXiv: 2005.07360

Convex Setting [1]

Assume you have $n = 3$ with $(x_i, y_i)_{i=1}^3 \to \{(\mathbf{e}_1, \beta_1^*), (\mathbf{e}_1, \beta_1^*), (\mathbf{e}_2, \beta_2^*)\}$

Empirical Loss

$$\hat{L}_n := \tfrac{2}{3}(\beta_1 - \beta_1^*)^2 + \tfrac{1}{3}(\beta_2 - \beta_2^*)^2$$

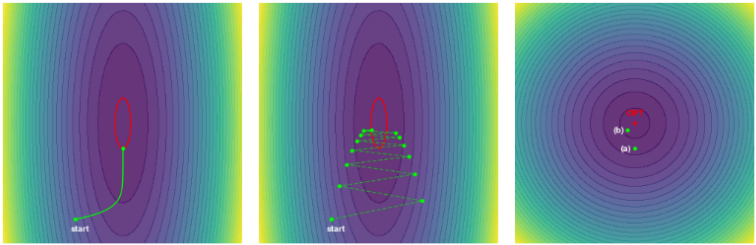
Population Loss

$$L_{\mathcal{D}} := \tfrac{1}{2}(\beta_1 - \beta_1^*)^2 + \tfrac{1}{2}(\beta_2 - \beta_2^*)^2$$

[1] P. Nakkiran (2020). "Learning Rate Annealing Can Provably Help Generalization, Even for Convex Problems". In: pp. 1–9. arXiv: 2005.07360

Convex Setting [1]

Start from 0-initialization and update with SGD until $\hat{L}_n = \varepsilon$ with small learning rate (a) and large learning rate with annealing (b)

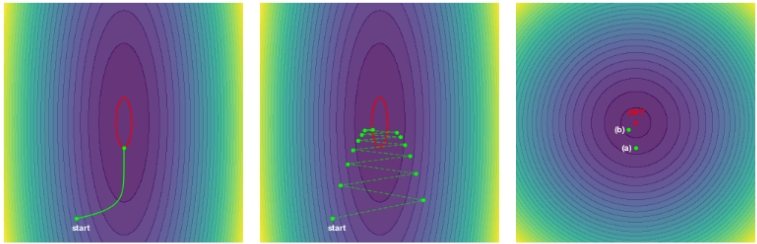


(a) Small learning rate. (b) Large, then small learning rate. (c) Test loss landscape.

[1] P. Nakkiran (2020). "Learning Rate Annealing Can Provably Help Generalization, Even for Convex Problems". In: pp. 1–9. arXiv: 2005.07360

Convex Setting [1]

Start from 0-initialization and update with SGD until $\hat{L}_n = \varepsilon$ with small learning rate (a) and large learning rate with annealing (b)



(a) Small learning rate. (b) Large, then small learning rate. (c) Test loss landscape.

So, large learning rate is regularizing the high curvature update and allow better generalization performance.

[1] P. Nakkiran (2020). "Learning Rate Annealing Can Provably Help Generalization, Even for Convex Problems". In: pp. 1–9. arXiv: 2005.07360

Conclusions

- Learning rate can have a regularizing effect in both convex and non-convex settings, although reasons are different.
- Investigation of learning rate remains an open question in more complex settings.