Assignment 11

112652010 韓韻宸

W1

Q: Mini-Batch SGD 的 batch size 要如何選擇?

- Gradient Noise Scale & 臨界批量：McCandlish et al., An Empirical Model of Large-Batch Training, 2018（OpenAI 解釋文 + 論文）。提出以梯度雜訊尺度預測「最大有用 batch size」。
  連結：OpenAI 解說頁、arXiv 論文

- 線性放大學習率 + warmup（大批量 ImageNet）：Goyal et al., Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, 2017。提出linear scaling rule 與 warmup，在 batch 高達 8192 時維持準確率。
  連結：arXiv PDF。

W3

Q: Why do the weights in the construction get so large when the required error ϵ is very small?

1. Lipschitz 與權重範數/結構

- Virmaux, A., & Scaman, K. (2018). Lipschitz regularity of deep neural networks: analysis and efficient estimation. NeurIPS.
  PDF：https://papers.neurips.cc/paper/7640-lipschitz-regularity-of-deep-neural-networks-analysis-and-efficient-estimation.pdf
  arXiv：https://arxiv.org/abs/1805.10965
- 深度 ReLU 近似率（ϵ\epsilonϵ 與模型複雜度/常數量化）
  Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. Neural Networks, 94, 103–114.
  arXiv：https://arxiv.org/abs/1610.01145
  Journal 版摘要：
  https://www.sciencedirect.com/science/article/abs/pii/S0893608017301545

W5

Q: Why are multivariate Gaussian contours ellipsoids determined by eigenvectors/eigenvalues of Σ?

- Reference: Johnson, R. A., & Wichern, D. W. (2002). Applied Multivariate Statistical Analysis (6th ed.), Ch. 4 — 等密度橢球的主軸與 Σ 的特徵向量/特徵值對應（掃描章節）。連結：Johnson & Wichern Ch.4 PDF。

Q: What breaks if Σ is not positive definite?

- Reference: Multivariate normal distribution（Wikipedia）— “Degenerate case” 小節：Σ 非滿秩/非正定時，對 Lebesgue 測度不具密度；需要用偽逆/偽行列式等處理。連結：Wikipedia 條目。

Q: Why does MLE use 1/n while the unbiased estimator uses 1/(n − 1)?

- **Reference:** *Bessel’s correction*（Wikipedia）— 解釋為何 MLE 取 1/n1/n1/n 有偏、而以 1/(n−1)1/(n-1)1/(n−1) 校正可得無偏。連結：Wikipedia 條目。

W7

Q: Why do we train on multiple noise levels instead of one fixed σ?

- **Reference:** Song, Y., & Ermon, S. (2019). *Generative Modeling by Estimating Gradients of the Data Distribution (Noise-Conditional Score Networks).* NeurIPS. 連結：論文 PDF。

Q: How is DSM related to denoising autoencoders?

- Reference: Vincent, P. (2011). A Connection Between Score Matching and Denoising Autoencoders. Neural Computation. 連結：期刊頁（摘要/全文）。

W8

Q: Why does score matching allow training without knowing the true data density p(x)?

- Hyvärinen, A. (2005). Estimation of Non-Normalized Statistical Models by Score Matching. JMLR, 6, 695–709.
  （提出以匹配「分數函數」∇_xlog $p(x)$ 的方式進行估計，目標可寫成只含模型可微項與邊界項，無需正規化常數與真實密度。）連結：JMLR PDF。

W10

Q: Can we derive the same probability flow ODE form for higher-dimensional SDEs, and what complications arise when the diffusion term g(x, t) becomes a matrix?

- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-Based Generative Modeling through Stochastic Differential Equations. ICLR — Appendix D derives the probability-flow ODE for the general multidimensional SDE with matrix diffusion, and discusses the resulting terms (e.g., involving $A=\sigma\sigma^{T}$) and their implications for likelihoods/sampling. 連結：arXiv（含附錄）