Problem Set 2

Physics 198: (Deep) Learning Mechanics

Covers Chapter 2: Toy Model I, Deep Linear Networks

Chapter 2 Exercises

These exercises fill in the calculations we deferred in the main text. None of them require any tools beyond what's in the chapter.

2.1 The loss in covariance form.

In the setup, we claimed that "with some matrix algebra" the mean squared error can be rewritten purely in terms of the data covariances. Verify this. Starting from

$$\mathcal{L}(W_1, \ldots, W_L) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\operatorname{Tr}(y - W_L \cdots W_1 x)(y - W_L \cdots W_1 x)^\top\Big],$$

expand the product inside the trace, use linearity of the trace and the expectation, and substitute the definitions $\Sigma_{xx} := \mathbb{E}_{\mathcal{D}}[xx^\top]$, $\Sigma_{yx} := \mathbb{E}_{\mathcal{D}}[yx^\top]$, and $\Sigma_{yy} := \mathbb{E}_{\mathcal{D}}[yy^\top]$ to arrive at

$$\mathcal{L}(W_1, \ldots, W_L) = \|\Sigma_{yx}\Sigma_{xx}^{-1/2} - W_L \cdots W_1 \Sigma_{xx}^{1/2}\|_{\mathrm{F}}^2 + \text{const.}$$

where const. $= \Sigma_{yy} - \Sigma_{yx}(\Sigma_{xx})^{-1}(\Sigma_{yx})^\top$. Confirm that the constant term does not depend on the weight matrices.

Hint: "complete the square" at the level of matrices—add and subtract $\Sigma_{yx}\Sigma_{xx}^{-1/2}$ inside the Frobenius norm.

2.2 The conservation law, from the symmetry.

The algebra trick used in the chapter to derive the conservation law $\frac{d}{dt}[W_2^\top W_2 - W_1 W_1^\top] = 0$ derives it as if by luck. This exercise re-derives it from the continuous symmetry of the loss, making clear *why* the trick works.

Throughout, use the matrix inner product $\langle A, B\rangle := \operatorname{Tr}\Big(A^\top B\Big)$, and recall the gradient flow $\dot{\theta} = -\nabla\mathcal{L}$ for $\theta = (W_1, W_2)$.

(a) *(Symmetry ⇒ orthogonality.)* Suppose $\mathcal{L}$ is invariant under a one-parameter family of transformations $\theta \mapsto \theta + \epsilon\, v(\theta)$, with generator (velocity field) v. Show that invariance, to first order in ϵ, is equivalent to

$$\langle v(\theta), \nabla\mathcal{L}(\theta)\rangle = 0 \quad \text{for all } \theta.$$

(b) *(Gradient flow never moves along the symmetry.)* Substitute the gradient flow equation to conclude that $\langle v(\theta), \dot{\theta}\rangle = 0$ along any trajectory. Interpret this geometrically: the gradient flow velocity is always orthogonal to the symmetry direction, so optimization never spends motion traveling along the symmetry orbit.

(c) *(The specific symmetry.)* For the deep linear network, the symmetry is $W_1 \mapsto AW_1,\ W_2 \mapsto W_2A^{-1}$. Take $A = e^{\epsilon M} \approx I + \epsilon M$ for an arbitrary generator matrix M, and show that the induced generator is $v = (MW_1,\ -W_2M)$.

(d) *(All generators at once.)* Plug this v into the orthogonality condition from (b), substitute $\nabla_{W_1}\mathcal{L} = -\dot{W}_1$ and $\nabla_{W_2}\mathcal{L} = -\dot{W}_2$, and use the cyclic property of the trace to rewrite both terms with $M^\top$ pulled to the front. Show the condition becomes

$$\mathrm{Tr}(M^\top \dot{W}_1 W_1^\top) = \mathrm{Tr}(M^\top W_2^\top \dot{W}_2).$$

(e) *(Conclude.)* Since this must hold for *every* generator M, deduce that $\dot{W}_1 W_1^\top = W_2^\top \dot{W}_2$, which is exactly $\frac{d}{dt}(W_2^\top W_2 - W_1 W_1^\top) = 0$. Note where the arbitrariness of M was essential: it is what upgrades a single scalar orthogonality condition into the full matrix-valued conservation law.

Remark. Compare this derivation to the algebra trick. The trick yields $\dot{W}_1 W_1^\top = W_2^\top \dot{W}_2$ by a fortunate cancellation; here the *same* equation appears as the direct consequence of "the flow is orthogonal to every symmetry direction." The luck of the trick was the symmetry in disguise all along.

2.3 Alignment collapses the model to its magnitudes.

Write the weight SVDs as $W_1 = U_1 S_1 V_1^\top$ and $W_2 = U_2 S_2 V_2^\top$, and recall the rotated model $\tilde{W}_2\tilde{W}_1 = U^{\star\top} W_2 W_1 V^\star$, where $\Sigma_{yx} = U^\star \Lambda^\star V^{\star\top}$. Show that under the three weight-alignment conditions,

$$U_2 = U^\star, \qquad V_1 = V^\star, \qquad V_2 = U_1,$$

the rotated model reduces to the product of magnitude matrices, $\tilde{W}_2\tilde{W}_1 = S_2 S_1$, with all rotation factors canceling. Identify which orthogonality relation kills each rotation factor.

2.4 Solving the scalar learning dynamics.

This is the climactic integration of the chapter. The μ-th end-to-end singular value obeys the separable ODE

$$\frac{d}{dt} s_\mu^2 = 2 s_\mu^2 (\lambda_\mu^\star - s_\mu^2).$$

Let $u := s_\mu^2$ for brevity, so that $\dot{u} = 2u(\lambda_\mu^\star - u)$.

(a) Separate variables and use the partial-fraction identity

$$\frac{1}{u(\lambda_\mu^\star - u)} = \frac{1}{\lambda_\mu^\star}\left(\frac{1}{u} + \frac{1}{\lambda_\mu^\star - u}\right)$$

to integrate both sides from 0 to t.

(b) Solve the resulting expression for $u = s_\mu^2(t)$ and confirm you recover the logistic solution

$$s_\mu^2(t) = \frac{\lambda_\mu^\star}{1 + \left(\frac{\lambda_\mu^\star}{s_\mu^2(0)} - 1\right) e^{-2\lambda_\mu^\star t}}.$$

(c) Check the two limits: confirm that $s_\mu^2(t) \to s_\mu^2(0)$ as $t \to 0$ and $s_\mu^2(t) \to \lambda_\mu^\star$ as $t \to \infty$.

2.5 Time-to-learn.

Using the logistic solution from the previous exercise, derive the halfway-time τ_μ at which mode μ reaches half its final strength, $s_\mu^2(\tau_\mu) = \frac{1}{2}\lambda_\mu^\star$. Show that, dropping order-one constants,

$$\tau_\mu \approx \frac{1}{\lambda_\mu^\star} \ln \frac{\lambda_\mu^\star}{s_\mu^2(0)}.$$

Then argue from this expression that (a) larger modes are learned first, and (b) the time-separation between consecutive modes widens as the initialization scale $s_\mu^2(0)$ shrinks—the mechanism behind the cleanly separated stairsteps in the loss curve.