Meeting Notes Pivotal

Guillaume, Alex Strang, Alex G.O.

June 23, 2025

1 Meeting 23-06-2025

- Discussed different assumptions behind SDEs for SGD and studying them in the saddle to saddle dynamics of linear networks.
- Strang shared takes on rotational of SGD noise (solenoidal flux?) vs noise anisotropy vs heavy tailness of noise
- Studying stationary distribution of SDE might not be that interesting because we are in a transient regime, far from time necessary for SGD to mix into the stationary distribution. We could measure this effect with area production rate statistics. There was something about production rate statistics as well.

One potential pushback in favour of stationary distribution being potentially useful is if we think that SGD can reach mixing time along subspaces of the parameter space. For example, while SGD might be always making progress by going down the loss landscape, many directions are degenerates and along these directions, SGD might have time to mix into some (projected) stationary distribution.

We discussed the different effective potential from the stationary distribution that one investigate depending on the temperature of the system. Consider the following SDE:

$$dX(t) = -\nabla L(X, w)dt + \sqrt{\beta D}dW_t \tag{1}$$

where $V(X_t)$ is the potential and β is the inverse temperature of the system.

This leads to the Fokker-Planck equation:

$$\frac{\partial p(X,t)}{\partial t} = \nabla \cdot (p(X,t)\nabla V(X)) + \nabla^2 \cdot \frac{D}{2} p(X,t) \tag{2}$$

We discussed the different effective potential from the stationary distribution that one investigate depending on the temperature of the system. For the stationary distribution $\pi(X)$ the FP equation becomes:

$$\nabla \cdot (\pi(X)\nabla V(X)) + \nabla^2 \cdot \frac{D}{2}\pi(X) = 0 \tag{3}$$

By letting

$$S_{\text{eff}} = -2\beta \log(\pi(X))$$

be an effective potential, we can rewrite the stationary FP equation as:

$$\beta\nabla(\nabla S_{\text{eff}} - \nu(X)) + \nabla S_{\text{eff}}(\nabla S_{\text{eff}} - \nu(X)) = 0 \tag{4}$$

with $\nu(X) = D^{-1}u(X)$ assuming a non degenerate loss landscape.

We can now distinguish different potentials depending on the temperature of the system:

- The loss
- The effective potential $S_{\text{eff}} = -2\beta \log(\pi(X))$
- The Helmoltz potential satisfying: $\beta\nabla(\nabla S_{\text{eff}} - \nu(X)) = 0$
- The quasi potential satisfying: $\nabla S_{\text{eff}}(\nabla S_{\text{eff}} - \nu(X)) = 0$

Note: we have to distinguish between the stationary distribution and the equilibrium distribution. It is possible that rotational current from non equilibrium distribution takes too long to kick in because of transient dynamics faster than the mixing time.

However, even if we don't care about the mixing time, the anisotropy of the noise is still an important effect: it changes the shape of the probability distribution in a way that is counterintuitive to the Langevin SDE picture since noise is dampened along degenerate directions, where it would have been easier to diffuse in the Langevin SDE picture.

Next steps: Two broad directions seem particularly salient to me. One is to check wether the stationary distribution could still be useful on subspaces of the training process within which SGD might hae time to mix into the stationary distribution and where non equilibrium stationary distribution is relevant to the dynamics. Another direction is to put this non equilibrium stationary distribution aside and focus on the anisotropy and heavy tailness of the noise. Intuitively, the latter seems simpler to look into and perhaps more important. My suggestion would be to look into the influence of noise anisotropy and heavy tailness on the saddle to saddle dynamics of linear networks and compare with Langevin SDE to show some important differences between the Boltzmann distribution obtained from Langevin SDE and this refined SDE model.