Stochastic Gradient Descent (SGD)

Heat Capacity with
Loss as Energy

and Learning Rate as Temperature
MCMC models of SGD

These Plots are from the original ResNet paper. Left plot is for CNNs without residual skip connections, the right plot is ResNet.

Thin lines are training error, thick lines are validation error.

In all cases $\eta$ is reduced twice, each time by a factor of 2.
Converged Loss as a Function of $\eta$

For each value of $\eta$ we converge at a loss $\mathcal{L}(\eta)$.

\[
\mathcal{L}(0) \doteq \lim_{\eta \to 0} \mathcal{L}(\eta) = \mathcal{L}(\Phi^*) \quad \Phi^* \text{ a local optimum}
\]

Can we do a Taylor expansion of $\mathcal{L}(\eta)$?

\[
\mathcal{L}(\eta) = \mathcal{L}(\Phi^*) + \left( \frac{d\mathcal{L}}{d\eta} \bigg|_{\eta=0} \right) \eta + \ldots
\]
Heat Capacity

\[ \mathcal{L}(\eta) = \mathcal{L}(\Phi^*) + \left( \frac{d\mathcal{L}}{d\eta} \bigg|_{\eta=0} \right) \eta + \ldots \]

Let \( b \) index a training example and let \( g_b \) denote \( \nabla_{\Phi} \mathcal{L}_b(\Phi) \) at \( \Phi = \Phi^* \).

Heat Capacity Theorem:

\[ \frac{\partial \mathcal{L}(\eta)}{\partial \eta} \bigg|_{\eta=0} = \frac{1}{4} E_b \|g_b\|^2 \]
Proof Step 1
Let $b$ index a training example and let $\mathcal{L}_b(\Phi^* + \Delta\Phi)$ be the loss on training example $b$ with model parameters $\Phi^* + \Delta\Phi$. We take a second order Taylor expansion.

\[
\mathcal{L}(\Phi) = E_b \mathcal{L}_b(\Phi)
\]

\[
\mathcal{L}_b(\Phi^* + \Delta\Phi) = \mathcal{L}_b(\Phi^*) + g_b \Delta\Phi + \frac{1}{2} \Delta\Phi^\top H_b \Delta\Phi
\]

\[
E_b \ g_b = 0
\]

\[
E_b \ H_b \quad \text{is positive definite}
\]
Proof: Step 2

Let $Q_\eta$ be the stationary distribution on $\Phi$ defined by the SGD stochastic process.

Let $P_\eta$ be the distribution on $\Delta \Phi = \Phi - \Phi^*$ with $\Phi \sim Q_\eta$.

\[
\mathcal{L}(\eta) = E_{\Delta \Phi \sim P_\eta} E_b \mathcal{L}_b + g_b \Delta \Phi + \frac{1}{2} \Delta \Phi^\top H_b \Delta \Phi
\]

\[
= E_b \mathcal{L}_b(\Phi^*) + E_{\Delta \Phi \sim P_\eta} (E_b g_b) \Delta \Phi + \frac{1}{2} \Delta \Phi^\top (E_b H_b) \Delta \Phi
\]

\[
= \mathcal{L}(\Phi^*) + E_{\Delta \Phi \sim P_\eta} \frac{1}{2} \Delta \Phi^\top (E_b H_b) \Delta \Phi
\]
Proof: Step 3

Because $P_\eta$ is a stationary distribution on $\Delta \Phi$ we must have

$$E_{\Delta \Phi \sim P_\eta} E_b \| \Delta \Phi - \eta (g_b + H_b \Delta \Phi) \|^2 = E_{\Delta \Phi \sim P_\eta} \| \Delta \Phi \|^2$$

$$E_{\Delta \Phi \sim P_\eta} E_b - 2\eta \Delta \Phi^\top (g_b + H_b \Delta \Phi) + \eta^2 \| (g_b + H_b \Delta \Phi) \|^2 = 0$$

$$E_{\Delta \Phi \sim P_\eta} \frac{1}{2} \Delta \Phi^\top (E_b \ H_b) \Delta \Phi) = \frac{\eta}{4} E_{\Delta \Phi \sim P_\eta} E_b \| (g_b + H_b \Delta \Phi) \|^2$$
Proof: Step 4

\[ \mathcal{L}(\eta) = \mathcal{L}(\Phi^*) + E_{\Delta \Phi \sim P_\eta} \frac{1}{2} \Delta \Phi^\top (E_b \, H_b) \Delta \Phi \]

\[ E_{\Delta \Phi \sim P_\eta} \frac{1}{2} \Delta \Phi^\top (E_b \, H_b) \Delta \Phi = \frac{\eta}{4} E_{\Delta \Phi \sim P_\eta} E_b \| (g_b + H_b \Delta \Phi) \|^2 \]

\[ \mathcal{L}(\eta) = \mathcal{L}(\Phi^*) + \frac{\eta}{4} E_{\Delta \Phi \sim P_\eta} E_b \| (g_b + H_b \Delta \Phi) \|^2 \]
Proof Step 5

\[ \mathcal{L}(\eta) = \mathcal{L}(\Phi^*) + \frac{\eta}{4} E_{\Delta \Phi \sim P_\eta} E_b \| (g_b + H_b \Delta \Phi) \|^2 \]

\[ \frac{\partial \mathcal{L}(\eta)}{\partial \eta} \bigg|_{\eta=0} = \frac{1}{4} \lim_{\eta \to 0} E_{\Delta \Phi \sim P_\eta} E_b \| (g_b + H_b \Delta \Phi) \|^2 \]

\[ = \frac{1}{4} E_b \| g_b \|^2 \]
END