1 Continuous Setting

Consider the traditional classification task where $x \in \mathbb{R}^d, y \in \mathbb{R}$. The goal is to find a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, such that:

$$\min_{f} Q(f) = L(f) + R(f), L(f) = E_{x,y}[l(f(x)), y],$$

where $l(\cdot)$ is defined to be the loss function and $R$ is a regularization function. Similar to Kernel methods, consider the two-level network given below to represent $f$:

$$f(\omega, \rho, x) = \int_{\mathbb{R}^d} \sigma(\theta, x)\omega(\theta)\rho(\theta)d\theta$$ (1)

where $\sigma(\theta, x) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a known real-valued function, $\omega(\theta) : \mathbb{R}^d \rightarrow \mathbb{R}$ is a real value function of $\theta$, and $\rho(\theta)$ is a probability density over $\theta$. For regularizer, we use

$$R(\omega, \rho) = \lambda_1 R_1(\omega, \rho) + \lambda_2 R_2(\rho)$$

where

$$R_1(\omega, \rho) = \int r_1(\omega(\theta))\rho(\theta)d\theta, r_1(\omega) = |\omega|^2$$

$$R_2(\rho) = \int r_2(\theta)\rho(\theta)d\theta, r_2(\theta) = ||\theta||^2$$

Next, we show a discrete NN approximates the continuous one when hidden nodes go to infinity and then drive the evolution rule of $\rho(\theta)$ and $\omega(\theta)$ from the (noisy) GD algorithm when the step size becomes small.

2 Discrete Setting

Consider a finite NN with the following form to approximate $f(\omega, \rho, x)$:

$$\hat{f}(\mu, \Theta, x) = \frac{1}{m} \sum_{j=1}^m \mu^j \sigma(\theta^j, x)$$ (2)

where $\Theta = \{\theta^j\}_{j=1}^m$.

The regularization terms are:

$$\hat{R}_1(\mu, \Theta) = \frac{1}{m} \sum_{j=1}^m r_1(\mu^j), \hat{R}_2(\Theta) = \frac{1}{m} \sum_{j=1}^m r_2(\theta^j).$$ (3)
and the training objective is:

\[
\hat{Q}(\mu, \Theta) = \mathbb{E}_{x,y} l(\hat{f}(\mu, \Theta, x), y) + \lambda_1 \hat{R}_1(\mu, \Theta) + \lambda_2 \hat{R}_2(\Theta). \tag{4}
\]

We can solve it through the standard (noisy) GD, the algorithm is given by:

**Step 0.** Initialize \(\mu_0 \sim P_{\mu,0}(\mu), \theta_0^j \sim P_{\theta,0}(\theta)\)

**Step 1.** Update \(\theta^j\) by

\[
\theta^j_{t+\Delta t} = \theta^j_t - \Delta t \nabla_{\theta^j} \left[ \hat{Q}(\mu_t, \Theta_t) \right] - \sqrt{\lambda_3} \xi^j_{t+1},
\]

where \(\Delta t\) is the step size and \(\xi^j_{t+1} \sim N\left(0, \sqrt{2\Delta t I_d}\right)\).

**Step 2.** Update \(\mu^j\) by

\[
\mu^j_{t+\Delta t} = \mu^j_t - \Delta t \nabla_{\mu^j} \left[ \hat{Q}(\mu_t, \Theta_t) \right] - \sqrt{\lambda_3} \zeta^j_{t+1},
\]

where \(\zeta^j_{t+1} \sim N(0, \sqrt{2\Delta t})\).

### 2.1 Plain GD

We first consider the unnoisy setting where \(\lambda_3 = 0\). We have the following Lemma.

**Lemma 1.** Suppose \(\lambda_3 = 0\). Suppose at time \(t \geq 0\), we have \(\theta^j_t \sim \rho_t\), and suppose \(\mu^j_t = \omega_t (\theta^j_t)\). Assume \(l'\) is continuous and \(\sigma\) is twice differentiable. For all \(x\), we have:

\[
\lim_{m \to \infty} \hat{f}(\mu_t, \Theta_t, x) = f(\omega_t, \rho_t, x) \tag{5}
\]

Furthermore, when \(\Delta t \to 0, m \to \infty\), we can derive,

\[
\frac{d\rho_t(\theta)}{dt} = -\nabla_{\theta} \cdot [\rho_t(\theta) g_2(t, \theta, \omega_t(\theta))]
\]

\[
\frac{d\omega_t(\theta)}{dt} = g_1(t, \theta, \omega_t(\theta)) - \nabla_{\theta} [\omega_t(\theta)] g_2(t, \theta, \omega_t(\theta)),
\]

where \(\nabla_{\theta}\cdot\) means the divergence, \(g_1\) and \(g_2\) satisfy:

\[
g_1(t, \theta, v) = -\mathbb{E}_{x,y} [l' (f(\omega_t, \rho_t, x), y) \sigma(\theta, x)] - \lambda_1 \nabla_v [r_1(v)]
\]

\[
g_2(t, \theta, v) = -\mathbb{E}_{x,y} [l' (f(\omega_t, \rho_t, x), y) v \nabla_{\theta}\sigma(\theta, x)] - \lambda_2 \nabla_{\theta} [r_2(\theta)].
\]

To prove the lemma, we utilize the tool with Fokker-Planck Equation to compute the evolution.
Background with Fokker-Planck Equation Suppose the movement of a particle in $m$-dimensional space can be characterized by the stochastic differential equation given below:

$$dx_t = g(x_t, t) dt + \sqrt{2\beta^{-1}\Sigma} dB_t$$

Let $x_t \sim p(x, t)$, the evolution of $p(x, t)$ is given by:

$$\frac{\partial p(x, t)}{\partial t} = \frac{\Sigma \Sigma^\top}{\beta} \nabla^2 p(x, t) - \nabla \cdot [p(x, t)g(x_t, t)]$$

Proof of Lemma 1. Let the $p_t(\theta, v)$ as the joint distribution for $(\theta, v)$:

$$(\theta^j_t, \mu^j_t) \sim p_t(\theta, v) = \rho_t \delta(v = \omega_t(\theta))$$

We can rewrite $f(\omega_t, \rho_t, x)$ as:

$$f(\omega_t, \rho_t, x) = \int_{\mathbb{R}^{d+1}} \sigma(\theta, x)p_t(\theta, v)d\theta dv.$$ 

By the Law of the Large number, when $m \to \infty$,

$$\hat{f}(\mu_t, \Theta_t, x) \to f(\omega_t, \rho_t, x).$$

Now we denote

$$\tilde{g}_2(t, \theta, v) = -\mathbb{E}_{x,y} \left[l' \left( \hat{f}(\mu_t, \Theta_t, x), y \right) v \nabla_\theta \sigma(\theta, x) \right] - \lambda_2 \nabla_\theta [r_2(\theta)]$$

From the update rule of GD, we have $\theta^j_{t+1} = \theta^j_t + \Delta t \tilde{g}_2(t, \theta^j_t, \mu^j_t)$. Let $\Delta t \to 0$, using $\mu^j_t = \omega_t(\theta^j_t)$, we have

$$\frac{d\theta^j_t}{dt} = \tilde{g}_2(t, \theta^j_t, \omega_t(\theta^j_t)).$$

By applying Fokker-Planck equation,

$$\frac{d\rho_t(\theta)}{dt} = -\nabla_\theta \cdot [\rho_t(\theta)\tilde{g}_2(t, \theta, \omega_t(\theta))]$$

As $m \to \infty$, and because $l'$ is continuous, $\sigma(\theta, x)$ and $\rho_t$ are also second-order smooth, we obtain:

$$\nabla_\theta \cdot [\rho_t(\theta)\tilde{g}_2(t, \theta, \omega_t(\theta))] - \nabla_\theta \cdot [\rho_t(\theta)g_2(t, \theta, \omega_t(\theta))] \overset{a.s.}{\to} 0$$

To prove the evolution form for $\omega_t(\theta)$, we let:

$$\tilde{g}_1(t, \theta, v) = -\mathbb{E}_{x,y} \left[l' \left( \hat{f}(\mu_t, \Theta_t, x), y \right) \sigma(\theta, x) \right] - \lambda_1 \nabla_x r_1(v).$$
Then, (ignoring the superscript $j$ since all $j$ have the same calculation)

$$\omega_{t+\Delta t} (\theta_{t+\Delta t})$$

$$=\omega_t (\theta_{t+\Delta t}) + \frac{d\omega_t (\theta_{t+\Delta t})}{dt} \Delta t + o(\Delta t)$$

$$=\omega_t (\theta_t + \tilde{g}_2 (t, \theta_t, \omega_t (\theta_t)) \Delta t + o(\Delta t)) + \frac{d\omega_t (\theta_{t+\Delta t})}{dt} \Delta t + o(\Delta t)$$

$$=\omega_t (\theta_t) + [\nabla_{\theta} \omega_t (\theta_t)] \cdot \tilde{g}_2 (t, \theta_t, \omega_t (\theta_t)) \Delta t + \frac{d\omega_t (\theta_{t+\Delta t})}{dt} \Delta t + o(\Delta t)$$

By the update rule $\omega_{t+\Delta t} (\theta_{t+\Delta t}) = \omega_t (\theta_t) + \tilde{g}_1 (t, \theta_t, \omega_t (\theta)) \Delta t$, we have:

$$\lim_{\Delta t \to 0} \frac{d\omega_t (\theta_{t+\Delta t})}{dt} = \tilde{g}_1 (t, \theta_t, \omega_t (\theta)).$$

The proof is finished by Let $\Delta t \to 0$, and let $m \to \infty$.

References