A Necessary Condition and Elimination of Local Minima for Deep Neural Networks

Appendix

A Proofs of Theorem 3

Let $z \in \mathbb{R}^{d_z}$ be an arbitrary local minimum of Q. From the convexity and differentiability of Q_i and R_i, we have that

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m} Q_i(\phi_i^z(\alpha,\epsilon,S)) + R_i(\varphi_i^z(\alpha,\epsilon,S)) \\
&\geq \frac{1}{m}\sum_{i=1}^{m} Q_i(\phi_i(z)) + R_i(\varphi_i(z)) + \partial Q_i(\phi_i(z))(\phi_i^z(\alpha,\epsilon,S)-\phi_i(z)) + \partial R_i(\varphi_i(z))(\varphi_i^z(\alpha,\epsilon,S)-\varphi_i(z)) \\
&= Q(z) + \frac{1}{m}\sum_{i=1}^{m} \partial Q_i(\phi_i(z))\phi_i^z(\alpha,\epsilon,S) + \partial R_i(\varphi_i(z))\varphi_i^z(\alpha,\epsilon,S) - \partial Q_i(\phi_i(z))\phi_i(z) - \partial R_i(\varphi_i(z))\varphi_i(z).
\end{aligned} \tag{5}$$

Since z is a local minimum of Q, by the definition of a local minimum, there exists $\epsilon_1 > 0$ such that $Q(z) \leq Q(z')$ for all $z' \in B(z,\epsilon_1)$. Then, for any $\epsilon \in [0,\epsilon_1/2)$ and any $\nu \in \mathcal{V}[z,\epsilon]$, the vector $(z+\epsilon v)$ is also a local minimum because

$$Q(z+\epsilon v) = Q(z) \leq Q(z'),$$

for all $z' \in B(z+\epsilon v, \epsilon_1/2) \subseteq B(z,\epsilon_1)$, where the set inclusion follows from the triangle inequality. This satisfies the definition of a local minimum for $(z+\epsilon v)$. Since the composition and the sums of differentiable functions are differentiable, the vector $(z+\epsilon v)$ is a differentiable local minimum. Therefore, from the first-order necessary condition of differentiable local minima, there exists $\epsilon_0 > 0$ such that for any $\epsilon \in [0,\epsilon_0)$, any $v \in \mathcal{V}[\theta,\epsilon]$, and any $k \in \{1,\dots,d_\theta\}$,

$$\partial_k Q(z+\epsilon v) = \frac{1}{m}\sum_{i=1}^{m} \partial Q_i(\phi_i(\theta))\partial_k\phi_i(z+\epsilon v) + \partial R_i(\varphi_i(\theta))\partial_k\varphi_i(z+\epsilon v) = 0, \tag{6}$$

where we used the fact that $\phi_i(z) = \phi_i(z+\epsilon v)$ and $\varphi_i(z) = \varphi_i(z+\epsilon v)$ for any $v \in \mathcal{V}[z,\epsilon]$. From (6), there exists $\epsilon_0 > 0$ such that for any $\epsilon \in [0,\epsilon_0)$, any $S \subseteq_{\text{fin}} \mathcal{V}[\theta,\epsilon]$ and any $\alpha \in \mathbb{R}^{d_\theta \times |S|}$,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m} \partial Q_i(\phi_i(z))\phi_i^z(\alpha,\epsilon,S) + \partial R_i(\varphi_i(z))\varphi_i^z(\alpha,\epsilon,S) \\
&= \sum_{k=1}^{d_z}\sum_{j=1}^{|S|} \alpha_{k,j}\left(\frac{1}{m}\sum_{i=1}^{m} \partial Q_i(\phi_i(z))\partial_k\phi_i(z+\epsilon S_j) + \partial R_i(\varphi_i(z))\partial_k\varphi_i(z+\epsilon S_j)\right) \\
&= 0
\end{aligned} \tag{7}$$

where the second line follows the definition of $\phi_i^z(\alpha,\epsilon,S)$ and $\varphi_i^z(\alpha,\epsilon,S)$, and the last line follows (6).

Furthermore,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m} \partial Q_i(\phi_i(z))\phi_i(z) + \partial R_i(\varphi_i(z))\varphi_i(z) \pm (1/\rho)\partial R_i(\varphi_i(z))\varphi_i(z) \\
&= \sum_{k=1}^{d_z} h(z)_k \left(\frac{1}{m}\sum_{i=1}^{m} \partial Q_i(\phi_i(z))\partial_k\phi_i(z) + \partial R_i(\varphi_i(z))\partial_k\varphi_i(z)\right) + (1-1/\rho)\frac{1}{m}\sum_{i=1}^{m} \partial R_i(\varphi_i(z))\varphi_i(z) \\
&= (1-1/\rho)\frac{1}{m}\sum_{i=1}^{m} \partial R_i(\varphi_i(z))\varphi_i(z)
\end{aligned} \tag{8}$$

where the second line follows the assumption of the existence of a function h for writing $\phi_i(z)$ and $\varphi_i(z)$, and the last line follows (6).

Substituting (7) and (8) into (5), there exists $\epsilon_0 > 0$ such that for any $\epsilon \in [0, \epsilon_0)$, any $S \subseteq_{\text{fin}} \mathcal{V}[\theta, \epsilon]$ and any $\alpha \in \mathbb{R}^{d_\theta \times |S|}$,

$$\frac{1}{m}\sum_{i=1}^{m} Q_i(\phi_i^z(\alpha, \epsilon, S)) + R_i(\varphi_i^z(\alpha, \epsilon, S)) \geq Q(z) - (1 - 1/\rho)\frac{1}{m}\sum_{i=1}^{m} \partial R_i(\varphi_i(z))\varphi_i(z).$$

This proves the main statement of the theorem. In the case of $\rho = 1$, this shows that on the one hand, there exists $\epsilon_0 > 0$ such that for any $\epsilon \in [0, \epsilon_0)$, $Q(z) \leq \inf\{\frac{1}{m}\sum_{i=1}^m Q_i(\phi_i^z(\alpha, \epsilon, S)) + R_i(\varphi_i^z(\alpha, \epsilon, S)) : S \subseteq_{\text{fin}} \mathcal{V}[z, \epsilon], \alpha \in \mathbb{R}^{d_z \times |S|}\}$. On the other hand, since $\phi_i(z) = \sum_{k=1}^{d_z} h(z)_k \partial_k \phi_i(z)$ and $\varphi_i(z) = \rho \sum_{k=1}^{d_z} h(z)_k \partial_k \varphi_i(z)$ with $\rho = 1$, we have that $Q(z) \geq \inf\{\frac{1}{m}\sum_{i=1}^m Q_i(\phi_i^z(\alpha, \epsilon, S)) + R_i(\varphi_i^z(\alpha, \epsilon, S)) : S \subseteq_{\text{fin}} \mathcal{V}[z, \epsilon], \alpha \in \mathbb{R}^{d_z \times |S|}\}$. Combining these yields the desires statement for the equality in the case of $\rho = 1$.

□

B Proofs of eliminating local minima

A high level idea behind the proofs of Theorems 1 and 2 in this section (instead of the proof via the PGB necessary condition) follows the idea utilized by Kawaguchi (2016) for deep linear networks. That is, we first obtain possible candidate local minima $\tilde{\theta}$ via the first-order necessary condition (i.e., $\{(\theta, a, b, W) : a = 0\}$), and then consider small perturbations of those candidate local minima. From the definition of local minima, the value at a possible local minimum $\tilde{\theta}$ must be less than or equal to the value at any sufficiently small perturbations of the given local minimum $\tilde{\theta}$. This condition imposes strong constraints on those candidate local minima, and turns out to be sufficient to prove the desired result with appropriate perturbations and rearrangements, together with the interpolation result with polynomial or simply based on linear algebra (i.e., we can interpolate m' points via polynomial as the corresponding matrix has rank m').

In all the proofs of Theorems 1 and 2 (including the proof with the PGB necessary condition), we let θ be arbitrary so that we can prove the failure mode of eliminating the suboptimal local minima in the next section (Theorem 4) by reusing these proofs. Let $\ell_y(q) = \ell(q, y)$, and let $\nabla \ell_y(\varphi(q)) = (\nabla \ell_y)(\varphi(q))$ be the gradient $\nabla \ell_y$ evaluated at an output $\varphi(q)$ of a function φ.

B.1 Proof of Theorem 1 without the PGB necessary condition

Proof of Theorem 1. Let θ be fixed. Let (a, b, W) be a local minimum of $\tilde{L}|_\theta(a, b, W) := \tilde{L}(\theta, a, b, W)$. Let $\tilde{L}|_{(\theta,W)}(a, b) = \tilde{L}(\theta, a, b, W)$. Since $\ell_y : q \mapsto \ell(q, y)$ is assumed to be differentiable, $\tilde{L}|_{(\theta,W)}$ is also differentiable (since a sum of differentiable functions is differentiable, and a composition of differentiable functions is differentiable). From the definition of a stationary point of a differentiable function $\tilde{L}|_{(\theta,W)}$, for all $k \in \{1, 2, \dots, d_y\}$, $a_k \frac{\partial \tilde{L}(\theta,a,b,W)}{\partial a_k} = \frac{1}{m}\sum_{i=1}^m (\nabla \ell_{y_i}(f(x_i;\theta) + g(x_i; a, b, W)))_k a_k \exp(w_k^\top x + b_k) + 2\lambda a_k = \frac{\partial \tilde{L}(\theta,a,b,W)}{\partial b_k} + 2\lambda a_k^2 = 2\lambda a_k^2 = 0$, which implies that $a_k = 0$ for all $k \in \{1, 2, \dots, d_y\}$, since $2\lambda \neq 0$. Therefore, we have that

$$a = 0. \tag{9}$$

This yields $g(x; a, b, W) = 0$, and

$$\tilde{L}(\theta, a, b, W) = L(\theta).$$

We now consider perturbations of a local minimum (a, b, W) of $L|_\theta$ with $a = 0$. Note that, among other equivalent definitions, a function $h : \mathbb{R}^d \to \mathbb{R}$ is said to be differentiable at $q \in \mathbb{R}^d$ if there exist a vector $\nabla h(q)$ and a function $\varphi(q; \cdot)$ (with its domain being a deleted neighborhood of the origin $0 \in \mathbb{R}^d$) such that $\lim_{\Delta q \to 0} \varphi(q; \Delta q) = 0$, and

$$h(q + \Delta q) = h(q) + \nabla h(q)^\top \Delta q + \|\Delta q\| \varphi(q; \Delta q),$$

for any non-zero vector $\Delta q \in \mathbb{R}^d$ that is sufficiently close to $0 \in \mathbb{R}^d$ (e.g., see fundamental increment lemma and the definition of differentiability for multivariable functions). Thus, with sufficiently small perturbations $\Delta a \in \mathbb{R}^{d_y}$ and $\Delta W = \begin{bmatrix}\Delta w_1 & \Delta w_2 & \dots & \Delta w_{d_y}\end{bmatrix} \in \mathbb{R}^{d_x \times d_y}$, there exists a function φ such that

$$
\begin{aligned}
&\tilde{L}(\theta, a+\Delta a, b, W+\Delta W) \\
&= \frac{1}{m}\sum_{i=1}^{m} \ell_{y_i}(f(x_i;\theta)+\Delta g_i) + \lambda\|\Delta a\|_2^2 \\
&= \frac{1}{m}\sum_{i=1}^{m} \ell_{y_i}(f(x_i;\theta)) + \nabla \ell_{y_i}(f(x_i;\theta))^\top \Delta g_i + \|\Delta g_i\|_2 \varphi(f(x_i;\theta);\Delta g_i) + \lambda\|\Delta a\|_2^2,
\end{aligned}
$$

where $\lim_{\Delta q\to 0}\varphi(f(x_i;\theta);\Delta q)=0$ and $\Delta g_i = g(x_i;\Delta a, b, W+\Delta W))$. Here, the last line follows the definition of the differentiability of ℓ_{y_i}, since $g(x_i;\Delta a, b, W+\Delta W)_k = \Delta a_k \exp(w_k^\top x_i + \Delta w_k^\top x_i + b_k)$ is arbitrarily small with sufficiently small Δa_k and Δw_k.

Combining the above two equations, since (a,b,W) is a local minimum, we have that, for any sufficiently small Δa and Δw,

$$
\begin{aligned}
&\tilde{L}(\theta, a+\Delta a, b, W+\Delta W) - \tilde{L}(\theta, a, b, W) \\
&= \frac{1}{m}\sum_{i=1}^{m} \nabla \ell_{y_i}(f(x_i;\theta))^\top \Delta g_i + \frac{1}{m}\sum_{i=1}^{m} \|\Delta g_i\|_2 \varphi(f(x_i;\theta);\Delta g_i) + \lambda\|\Delta a\|_2^2 \\
&\geq 0.
\end{aligned}
$$

Rearranging with $\Delta a = \epsilon v$ such that $\epsilon > 0$ and $\|v\|_2 = 1$, and with $\Delta\tilde{g}_i = g(x_i; v, b, W+\Delta W)$,

$$
\frac{\epsilon}{m}\sum_{i=1}^{m} \nabla \ell_{y_i}(f(x_i;\theta))^\top \Delta\tilde{g}_i \geq -\frac{\epsilon}{m}\sum_{i=1}^{m} \|\Delta\tilde{g}_i\|_2 \varphi(f(x_i;\theta);\epsilon\Delta\tilde{g}_i) - \lambda\epsilon^2\|v\|_2^2,
$$

since $\Delta g_i = \epsilon\Delta\tilde{g}_i$. With $\epsilon > 0$, this implies that

$$
\frac{1}{m}\sum_{i=1}^{m} \nabla \ell_{y_i}(f(x_i;\theta))^\top \Delta\tilde{g}_i \geq -\frac{1}{m}\sum_{i=1}^{m} \|\Delta\tilde{g}_i\|_2 \varphi(f(x_i;\theta);\epsilon\Delta\tilde{g}_i) - \lambda\epsilon\|v\|_2^2.
$$

Since $\varphi(f(x_i;\theta);\epsilon\Delta\tilde{g}_i) \to 0$ and $\lambda\epsilon\|v\|_2^2 \to 0$ as $\epsilon \to 0$ ($\epsilon \neq 0$),

$$
\sum_{i=1}^{m} \nabla \ell_{y_i}(f(x_i;\theta))^\top g(x_i; v, b, W+\Delta W) \geq 0.
$$

For any $k \in \{1, 2, \dots, d_y\}$, by setting $v_{k'} = 0$ for all $k' \neq k$, we have that

$$
v_k \sum_{i=1}^{m} (\nabla \ell_{y_i}(f(x_i;\theta)))_k \exp(w_k^\top x_i + \Delta w_k^\top x_i + b_k) \geq 0,
$$

for any $v_k \in \mathbb{R}$ such that $|v_k| = 1$. With $v_k \in \{-1, 1\}$,

$$
\sum_{i=1}^{m} (\nabla \ell_{y_i}(f(x_i;\theta)))_k \exp(w_k^\top x_i + b_k)\exp(\Delta w_k^\top x_i) = 0.
$$

By setting $\Delta w_k = \bar{\epsilon}_k u_k$ such that $\bar{\epsilon}_k > 0$ and $\|u\|_2 = 1$,

$$
\sum_{t=0}^{\infty} \frac{\bar{\epsilon}_k^t}{t!} \sum_{i=1}^{m} (\nabla \ell_{y_i}(f(x_i;\theta)))_k \exp(w_k^\top x_i + b_k)(u_k^\top x_i)^t = 0,
$$

since $\exp(q) = \lim_{T\to\infty}\sum_{t=0}^{T}\frac{q^t}{t!}$ and a finite sum of limits of convergent sequences is the limit of the finite sum. Rewriting this using $z_t = \sum_{i=1}^{m} (\nabla \ell_{y_i}(f(x_i;\theta)))_k \exp(w_k^\top x_i + b_k)(u_k^\top x_i)^t$,

$$
\lim_{T\to\infty} \sum_{t=0}^{T} \frac{\bar{\epsilon}_k^t}{t!} z_t = 0. \tag{10}
$$

We now show that $z_p = 0$ for all $p \in \mathbb{N}_0$ by induction. Consider the base case with $p = 0$. Equation (10) implies that

$$\lim_{T\to\infty}\left(z_0 + \sum_{t=1}^{T} \frac{\bar\epsilon_k^t}{t!} z_t\right) = z_0 + \lim_{T\to\infty}\sum_{t=1}^{T}\frac{\bar\epsilon_k^t}{t!} z_t = 0$$

since $\lim_{T\to\infty}\sum_{t=1}^{T}\frac{\bar\epsilon_k^t}{t!} z_t$ exists (which follows that $\lim_{T\to\infty}\sum_{t=0}^{T}\frac{\bar\epsilon_k^t}{t!} z_t = 0$ exists). Here, $\lim_{T\to\infty}\sum_{t=1}^{T}\frac{\bar\epsilon_k^t}{t!} z_t \to 0$ as $\bar\epsilon \to 0$, and hence $z_0 = 0$. Consider the inductive step with the inductive hypothesis that $z_t = 0$ for all $t \le p - 1$. Similarly to the base case, Equation (10) implies

$$\sum_{t=0}^{p-1}\frac{\bar\epsilon_k^t}{t!} z_t + \frac{\bar\epsilon_k^p}{p!} z_p + \lim_{T\to\infty}\sum_{t=p+1}^{T}\frac{\bar\epsilon_k^t}{t!} z_t = 0.$$

Multiplying $p!/\bar\epsilon_k^p$ on both sides, since $\sum_{t=0}^{p-1}\frac{\bar\epsilon_k^t}{t!} z_t = 0$ from the inductive hypothesis,

$$z_p + \lim_{T\to\infty}\sum_{t=p+1}^{T}\frac{\bar\epsilon_k^{t-p} p!}{t!} z_t = 0.$$

Since $\lim_{T\to\infty}\sum_{t=p+1}^{T}\frac{\bar\epsilon_k^{t-p} p!}{t!} z_t \to 0$ as $\bar\epsilon \to 0$, we have that $z_p = 0$, which finishes the induction. Therefore, for any $k \in \{1, 2, \dots, d_y\}$ and any $p \in \mathbb{N}_0$,

$$\sum_{i=1}^{m} (\nabla \ell_{y_i}(f(x_i;\theta)))_k \exp(w_k^\top x_i + b_k)(u_k^\top x_i)^p = 0. \tag{11}$$

Let $x \otimes x$ be the tensor product of the vectors x and $x^{\otimes p} = x \otimes \cdots \otimes x$ where x appears p times. For a p-th order tensor $M \in \mathbb{R}^{d\times\cdots\times d}$ and p vectors $u^{(1)}, u^{(2)}, \dots, u^{(p)} \in \mathbb{R}^d$, defines

$$M(u_k^{(1)}, u_k^{(2)}, \dots, u_k^{(p)}) = \sum_{1\le i_1\cdots i_p\le d} M_{i_1\cdots i_p} u_{i_1}^{(1)} \cdots u_{i_p}^{(p)}.$$

Let $\xi_{i,k} = (\nabla \ell_{y_i}(f(x_i;\theta)))_k \exp(w_k^\top x_i + b_k)$. Then, for any $k \in \{1, 2, \dots, d_y\}$ and any $p \in \mathbb{N}_0$,

$$\begin{aligned} \max_{\substack{u^{(1)},\dots,u^{(p)}:\\ \|u^{(1)}\|_2=\cdots=\|u^{(p)}\|_2=1}} \left(\sum_{i=1}^{m}\xi_{i,k} x_i^{\otimes p}\right)(u^{(1)},\dots,u^{(p)}) &= \max_{u:\|u\|_2=1}\left(\sum_{i=1}^{m}\xi_{i,k} x_i^{\otimes p}\right)(u,u,\dots,u) \\ &= \max_{u:\|u\|_2=1}\sum_{i=1}^{m}\xi_{i,k}(u^\top x_i)^p = 0. \end{aligned}$$

where the first line follows theorem 2.1 in (Zhang et al., 2012), and the last line follows Equation (11). This implies that

$$\sum_{i=1}^{m} (\nabla \ell_{y_i}(f(x_i;\theta)))_k \exp(w_k^\top x_i + b_k)\operatorname{vec}(x_i^{\otimes p}) = 0 \in \mathbb{R}^{d_x^p}. \tag{12}$$

Using Equation (12), we now prove statement (i). For any θ', there exist p and $u_{t,k}$ (for $t = 0, \dots, p$ and $k = 1, \dots, d_y$) such that

$$\begin{aligned} m(L(\theta') - L(\theta)) &\ge \sum_{i=1}^{m} \nabla \ell_{y_i}(f(x_i;\theta))^\top (f(x_i;\theta') - f(x_i;\theta)) \\ &= \sum_{j=1}^{m'}\sum_{i\in\mathcal{I}_j} \nabla \ell_{y_i}(f(x_i;\theta))^\top (f(x_i;\theta') - f(x_i;\theta)) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{m'} \sum_{k=1}^{d_y} \underbrace{(f(\bar{x}_j;\theta') - f(\bar{x}_j;\theta))_k}_{=\exp(w_k^\top \bar{x}_j + b_k) \sum_{t=0}^{p} u_{t,k}^\top \operatorname{vec}(\bar{x}_j^{\otimes t})} \sum_{i \in \mathcal{I}_j} \nabla \ell_{y_i}(f(x_i;\theta))_k \\
&= \sum_{t=0}^{p} \sum_{k=1}^{d_y} u_{t,k}^\top \underbrace{\sum_{i=1}^{m} \nabla \ell_{y_i}(f(x_i;\theta))_k \exp(w_k^\top x_i + b_k) \operatorname{vec}(x_i^{\otimes t})}_{= 0 \text{ from Equation (12)}} \\
&= 0,
\end{aligned}
$$

where the first line follows from the assumption that ℓ_{y_i} is convex and differentiable, and the third line follows from the fact that $\bar{x}_j = x$ for all $x \in \mathcal{I}_j$. The forth line follows from the fact that the vector $\operatorname{vec}(x_i^{\otimes t})$ contains all monomials in x_i of degree t, and m' input points $\bar{x}_1, \dots, \bar{x}_{m'}$ are distinct, which allows the basic existence (and construction) result of a polynomial interpolation of the finite m' points; i.e., with p sufficiently large ($p = m' - 1$ is sufficient), for each k, there exists $u_{t,k}$ such that $\sum_{t=0}^{p} u_{t,k}^\top \operatorname{vec}(\bar{x}_j^{\otimes t}) = q_{j,k}$ for any $q_{j,k} \in \mathbb{R}$ for all $j \in \{1, \dots, m'\}$ (e.g., see equation (1.9) in Gasca and Sauer 2000), in particular, including $q_{j,k} = (f(\bar{x}_j;\theta') - f(\bar{x}_j;\theta))_k \exp(-w_k^\top \bar{x}_j - b_k)$.

Therefore, we have that, for any θ', $L(\theta') \geq L(\theta)$, which proves statement (i). Statement (ii) directly follows from Equation (9). □

B.2 Proof of Theorem 2

Proof of Theorem 2. Let θ be fixed. Let (a, b, W) be a local minimum of $\tilde{L}|_\theta(a, b, W) := \tilde{L}(\theta, a, b, W)$. Then, for any $k \in \{1, 2, \dots, d_y\}$, there exist p and $u_{t,k}$ (for $t = 0, \dots, p$) such that

$$
\begin{aligned}
\sum_{i=1}^{m} (\nabla \ell_{y_i}(f(x_i;\theta)))_k^2 &= \sum_{j=1}^{m'} |\mathcal{I}_j| (\nabla \ell_{f^*(\bar{x}_j)}(f(\bar{x}_j;\theta)))_k^2 \\
&= \sum_{t=0}^{p} u_{t,k}^\top \sum_{i=1}^{m} (\nabla \ell_{y_i}(f(x_i;\theta)))_k \exp(w_k^\top x_i + b_k) \operatorname{vec}(x_i^{\otimes t}) \\
&= 0,
\end{aligned}
$$

where the first line utilizes Assumption 3. The second line follows from the fact that since m' input points $\bar{x}_1, \dots, \bar{x}_{m'}$ are distinct, with p sufficiently large ($p = m' - 1$ is sufficient), for each k, there exist $u_{t,k}$ for $t = 0, \dots, p$ such that $\sum_{t=0}^{p} u_{t,k}^\top \operatorname{vec}(x_i^{\otimes t}) = (\nabla \ell_{f^*(\bar{x}_j)}(f(\bar{x}_j;\theta)))_k \exp(-w_k^\top \bar{x}_j - b_k) |\mathcal{I}_j|^{-1}$ (similarly to the proof of Theorem 1). The third line follows from Equation (12). Here, Equation (12) still holds since it is obtained in the proof of Theorem 1 under only the assumption that the function $\ell_{y_i} : q \mapsto \ell(q, y_i)$ is differentiable for any $i \in \{1, \dots, m\}$, which is still satisfied by Assumption 2.

This implies that for all $i \in \{1, \dots, m\}$, $\nabla \ell_{y_i}(f(x_i;\theta)) = 0$, which proves statement (iii) because of Assumption 2. Statement (i) directly follows from Statement (iii). Statement (ii) directly follows from Equation (9). □

C Proof of Theorem 4

The proofs of Theorems 1 and 2 (including the proof via the PGB necessary condition) are designed such that the proof of Theorem 4 is simple, as shown below. Given a function $\varphi(q) \in \mathbb{R}^d$ and a vector $v \in \mathbb{R}^{d'}$, let $\frac{\partial \varphi(q)}{\partial v}$ be a $d \times d'$ matrix with each entry $(\frac{\partial \varphi(q)}{\partial v})_{i,j} = \frac{\partial (\varphi(q))_i}{\partial v_j}$.

Proof of Theorem 4. Let Assumption 1 hold (instead of Assumptions 2 and 3). In the both versions of our proofs of Theorem 1, θ was arbitrary and (a, b, W) was an arbitrary local minimum of $\tilde{L}|_\theta(a, b, W) := \tilde{L}(\theta, a, b, W)$. Thus, the same proof proves that, for any θ, at every local minimum $(a, b, W) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_x \times d_y}$ of $\tilde{L}|_\theta$, θ is a global minimum of L. Thus, based on the logical equivalence ($p \to q \equiv \neg q \to \neg p$), if θ is a not global minimum of L, then there is no local minimum $(a, b, W) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_x \times d_y}$ of $\tilde{L}|_\theta$, proving the first statement in the case of using Assumption 1. Instead of Assumption 1, if Assumptions 2 and 3 hold, then the exact same proof as above (with Theorem 1 being replaced by Theorem 2) proves the first statement.

Example 1 with the square loss or the smoothed hinge loss suffices to prove the second statement. However, to obtain better theoretical insight, let us consider a more general construction of the desired tuples $(\ell, f, \{(x_i, y_i)\}_{i=1}^m)$ to prove the second statement. Let $\theta \in \mathbb{R}^{d_\theta}$. In addition, let $A[\theta] = \frac{1}{m}[(\frac{\partial f(x_1;\theta)}{\partial \theta})^\top \ \cdots \ (\frac{\partial f(x_m;\theta)}{\partial \theta})^\top] \in \mathbb{R}^{d_\theta \times (md_y)}$ be a matrix, and $r[\varphi] = [\nabla \ell_{y_1}(\varphi(x_1))^\top \ \cdots \ \nabla \ell_{y_m}(\varphi(x_m))^\top]^\top \in \mathbb{R}^{md_y}$ be a column vector given a function $\varphi : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$. Then,

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{1}{m}\sum_{i=1}^{m} \nabla \ell_{y_i}(f(x_i;\theta))^\top \frac{\partial f(x_i;\theta)}{\partial \theta} = (A[\theta] r[f(\cdot;\theta)])^\top,$$

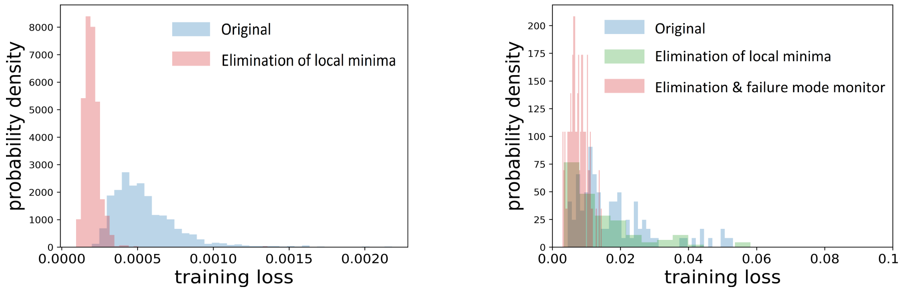
and

$$\frac{\partial \tilde{L}(\theta, a, b, W)}{\partial \theta} = (A[\theta] r[f(\cdot;\theta) + g(\cdot; a, b, W)])^\top.$$

Here, the equality $A[\theta]r[f(\cdot;\theta)] = 0$ is equivalent to $r[f(\cdot;\theta)] \in \mathrm{Null}(A[\theta])$, where $\mathrm{Null}(A[\theta])$ is the null space of the matrix $A[\theta]$. Therefore, any tuple $(\ell, f, \{(x_i, y_i)\}_{i=1}^m)$ such that $r[f(\cdot;\theta)] \in \mathrm{Null}(A[\theta]) \Rightarrow r[f(\cdot;\theta) + g(\cdot; a, b, W)] \in \mathrm{Null}(A[\theta])$ at a suboptimal θ suffices to provide a proof for the second statement. An (infinite) set of tuples $(\ell, f, \{(x_i, y_i)\}_{i=1}^m)$ such that there exists a suboptimal θ of L with $A[\theta] = 0$ (e.g., Example 1) satisfies this condition, which proves the second statement. □

D Additional numerical examples for good cases

For using $\tilde{L}$ instead of L, we show the failure mode and 'bad-case' scenarios in Section 6 and Appendix E. Accordingly, to have a balance, this section considers some of 'good-case' scenarios where using $\tilde{L}$ instead of L helps optimization of L. Figure 3 shows the histograms of training loss values after training with original networks f minimizing L, and modified networks $\tilde{f}$ minimizing $\tilde{L}$ with and without the failure mode detector based on Theorems 1, 2 and 4. We used a simple failure mode detector, which automatically restarted the optimizer to a random point during training when $\|a\|_2 + \|b\|_2 + \|W\|_2 \geq 7$. The histograms were plotted with the results of 1000 random trials for Semeion dataset and of 100 random trials for KMNIST dataset, for each method. Semeion (Brescia, 1994) is a dataset of handwritten digits and KMNIST (Clanuwat et al., 2018) is a dataset of Japanese letters. We used the exact same experimental settings for both the original networks f and the modified networks $\tilde{f}$ with and without the failure mode detector. We used a standard variant of LeNet (LeCun et al., 1998) with ReLU activations: two convolutional layers with 64 5×5 filters, followed by a fully-connected layer with 1024 output units and the output layer. The AdaGrad optimizer was employed with the mini-batch size of 64.



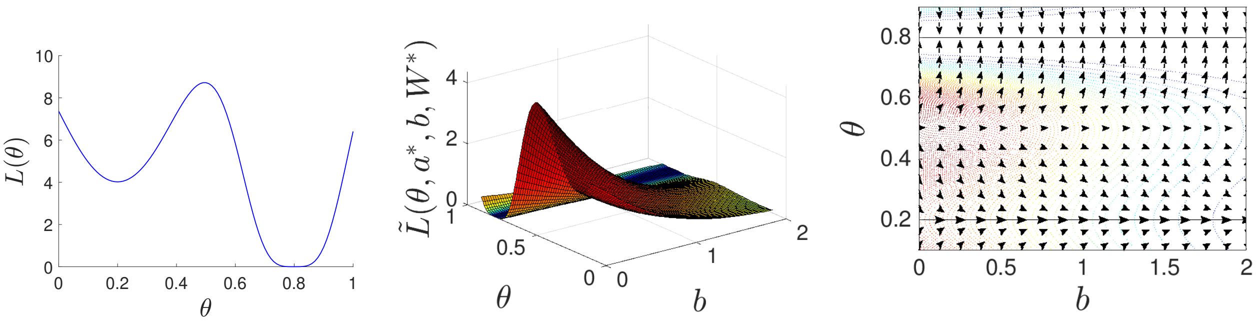
(a) Semeion

(b) KMNIST

Figure 3: Histogram of loss values after training with original networks f minimizing L (original), modified networks $\tilde{f}$ minimizing $\tilde{L}$ (elimination of local minima), and modified networks $\tilde{f}$ minimizing $\tilde{L}$ with the failure mode detector (elimination & failure mode monitor). The plotted training loss values are the values of the standard training objective L for both original networks f (minimizing L) and modified networks $\tilde{f}$ (minimizing $\tilde{L}$) with and without the failure mode detector. The elimination of local minima helped a gradient-based method for Semeion, and did not help it much for KMNIST. For KMNIST, the novel failure mode of the elimination was detected by monitoring the norms of (a, b, W) to restart and search a better subspace.

E Additional numerical and analytical examples to illustrate the failure mode

Figure 4 illustrates the novel failure mode proven by Theorem 4. The setting used for plotting Figure 4 is exactly same as that in Figure 1 (i.e., Example 1) except that $\ell(f(x_1;\theta),y_1) = (f(x_1;\theta) - y_1)^2$ and $y_1 = f(x_1; 0.8)$.



(a) original objective function L (b) modified objective function $\tilde{L}$ (c) negative gradient directions of $\tilde{L}$

Figure 4: Illustration of the failure mode suggested by Theorem 4 with the squared loss. The qualitatively identical behavior as that in Figure 1 can be observed.

Examples 6 and 7 illustrate the same phenomena as those in Examples 3 and 4 with a smoothed hinge loss instead of the squared loss.

Example 6. Let $m = 1$ and $d_y = 1$. In addition, $L(\theta) = \ell(f(x_1;\theta), y_1) = (\max(0, 1 - y_1 f(x_1;\theta))^3$. Accordingly, $\tilde{L}(\theta, a, b, W) = (\max(0, 1 - y_1 f(x_1;\theta) - y_1 a \exp(w^\top x_1 + b))^3 + \lambda a^2$. Let θ be a non-global minimum of L as $f(x_1;\theta) \neq y_1$, in particular, by setting $f(x_1;\theta) = -1$ and $y_1 = 1$. Then, $L(\theta) = 8$. If (a, b, W) is a local minimum, we must have $a = 0$ similarly to Example 3, yielding that $\tilde{L}(\theta, a, b, W) = 8$. However, a point with $a = 0$ is not a local minimum, since with $a > 0$ being sufficiently small,

$$\tilde{L}(\theta, a, b, W) = (2 - a\exp(w^\top x_1 + b))^3 + \lambda a^2 < 8.$$

Hence, there is no local minimum $(a, b, W) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{d_x}$ of $\tilde{L}|_\theta$. Indeed, if we set $a = -2\exp(-1/\epsilon)$ and $b = 1/\epsilon - w^\top x_1$, $\tilde{L}(\theta, a, b, W) = \lambda \exp(-2/\epsilon) \to 0$ as $\epsilon \to 0$, and hence as $a \to 0^-$ and $b \to \infty$. This illustrates the case in which (a, b) does not attain a solution in $\mathbb{R} \times \mathbb{R}$. The identical conclusion holds with the general case of $f(x_1;\theta) \neq y_1$ by following the same logic.

Example 7. Let $m = 2$ and $d_y = 1$. In addition, $L(\theta) = (\max(0, 1 - y_1 f(x_1;\theta))^3 + (\max(0, 1 - y_2 f(x_2;\theta))^3$. Moreover, let $x_1 \neq x_2$. Finally, let $f(x_1;\theta) = -1, f(x_2;\theta) = 1$, $y_1 = 1$, and $y_2 = -1$. If (a, b, W) is a local minimum, we must have $a = 0$ similarly to Example 3, yielding $\tilde{L}(\theta, a, b, W) = 16$. However, a point with $a = 0$ is not a local minimum, which follows from the perturbations of (a, W) in the same manner as in Example 4. Therefore, there is no local minimum (a, b, W) of $\tilde{L}|_\theta$. Indeed, if we set $a = 2\exp(-1/\epsilon)$, $b = 1/\epsilon - w^\top x_1$, and $w = -\frac{1}{\epsilon}(x_2 - x_1)$,

$$\tilde{L}(\theta, a, b, W) = (2 + 2\exp(-\|x_2 - x_1\|_2^2/\epsilon))^3 + \lambda\exp(-2/\epsilon) \to 8$$

as $\epsilon \to 0$, and hence as $a \to 0^-$, $b \to \infty$ and $\|w\| \to \infty$, illustrating the case in which (a, b, W) does not attain a solution in $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^{d_x}$.