A. Derivation of Metric Tensor and Update Direction

In this appendix we provide intuition for what $\tilde{l}$ does relative to l, and show how G_i^x encodes the dissimilarity function in (1). We begin by considering the stochastic gradient descent learning rule as an example to understand what the $\partial f(\cdot, \beta_i)/\partial \beta_i$ terms in a learning rule do, and how they should be changed to decouple the decisions of which learning rule to use and which parameterized function to use. For simplicity, in this appendix we consider the setting where $\beta_i \coloneqq \theta_{i-1}$ and we use the shorthand $\theta_i \coloneqq l_i(f, \theta_0, \omega)$. Also, recall that in all appendices we use the shorthands: $d \coloneqq d_i^{x,\theta_0,\omega}$, and $G_i^x \coloneqq G_i^x$.

The stochastic gradient descent update to make f approximate some target function, $f^\star$, can be written as:

$$\theta_i = \theta_{i-1} + \alpha_i \underbrace{\left(f^\star(X_{i-1}(\omega)) - f(X_{i-1}(\omega), \theta_{i-1})\right)}_{=:\delta_{i-1}} \frac{\partial f}{\partial \theta_{i-1}}(X_{i-1}(\omega), \theta_{i-1}), \tag{3}$$

where $X_{i-1} : \Omega \to \mathcal{X}$ is a random variable, and where $(\alpha_i)_{i=1}^\infty$ is a sequence of small positive real-valued step sizes. For brevity, hereafter we write X_{i-1} as shorthand for $X_{i-1}(\omega)$. In (3) the δ_{i-1} term is an *error term*. If δ_{i-1} is positive, it means that θ_i should be selected to make $f(X_{i-1}, \theta_i)$ larger than $f(X_{i-1}, \theta_{i-1})$. Similarly, if δ_{i-1} is negative, then it means that θ_i should be selected to make $f(X_{i-1}, \theta_i)$ smaller than $f(X_{i-1}, \theta_{i-1})$. This intuition is accomplished in (3) by multiplying δ_{i-1} by $\frac{\partial f}{\partial \theta_{i-1}}(X_{i-1}, \theta_{i-1})$, which is a direction of change to θ_{i-1} that increases the value of $f(X_{i-1}, \theta_{i-1})$.

However, there are many directions, Δ_{i-1}, of change to the parameters, θ_{i-1}, that would cause $f(X_{i-1}, \theta_{i-1})$ to increase. In general, we could change the learning rule to be:

$$\theta_i = \theta_{i-1} + \delta_{i-1}\Delta_{i-1},$$

for any Δ_{i-1} such that (for infinitesimal α_i) $f(X_{i-1}, \theta_{i-1} + \alpha_i\Delta_{i-1}) \geq f(X_{i-1}, \theta_{i-1})$. However, some directions, Δ_{i-1}, are "better" than others. The error term, δ_{i-1}, describes whether $f(X_{i-1}, \theta_{i-1})$ should be bigger or smaller, but does not describe whether $f(x, \theta_{i-1})$ should be bigger or smaller for any $x \neq X_{i-1}$. Some directions, Δ_{i-1}, might cause $f(X_{i-1}, \theta_{i-1} + \alpha_i\Delta_{i-1})$ to increase slowly as α_i increases, but $f(x, \theta_{i-1} + \alpha_i\Delta_{i-1})$ to increase or decrease quickly as α_i increases, for some $x \neq X_{i-1}$. These Δ_{i-1} are not desirable because δ_{i-1} does not describe whether $f(x, \theta_{i-1})$ should be bigger or smaller. We desire a direction, Δ_{i-1}, that does the opposite: it should cause $f(X_{i-1}, \theta_{i-1} + \alpha_i\Delta_{i-1})$ to increase quickly with α_i, and $f(x, \theta_{i-1} + \alpha_i\Delta_{i-1})$ to change slowly with α_i for all $x \neq X_{i-1}$.

We will focus our attention of the first constraint: we will find a direction, Δ_{i-1}, that causes $f(X_{i-1}, \theta_{i-1} + \alpha_i\Delta_{i-1})$ to increase *as quickly as possible* with α_i. That is, we will select Δ_{i-1} to be a direction (vector of length one) such that for a step of infinitesimal length, α_i, $f(X_{i-1}, \theta_i + \alpha_i\Delta_{i-1})$ is maximized. More formally, we will select

$$\begin{aligned}
\Delta_{i-1} &\coloneqq \lim_{\alpha_i \to 0} \operatorname*{arg\,max}_{\Delta_{i-1} \in \{\Delta \in \mathbb{R}^n : \|\Delta\| = 1\}} f(X_{i-1}, \theta_{i-1} + \alpha_i\Delta_{i-1}) \\
&\overset{\text{(a)}}{=} \lim_{\alpha_i \to 0} \operatorname*{arg\,max}_{\Delta_{i-1} \in \{\Delta \in \mathbb{R}^n : \|\Delta\| = 1\}} \frac{\partial f}{\partial \theta_{i-1}}(X_{i-1}, \theta_{i-1})^\intercal(\alpha_i\Delta_{i-1}) + O(\alpha_i^2) \\
&= \lim_{\alpha_i \to 0} \operatorname*{arg\,max}_{\Delta_{i-1} \in \{\Delta \in \mathbb{R}^n : \|\Delta\| = 1\}} \frac{\partial f}{\partial \theta_{i-1}}(X_{i-1}, \theta_{i-1})^\intercal(\alpha_i\Delta_{i-1}) \\
&= \operatorname*{arg\,max}_{\Delta_{i-1} \in \{\Delta \in \mathbb{R}^n : \|\Delta\| = 1\}} \frac{\partial f}{\partial \theta_{i-1}}(X_{i-1}, \theta_{i-1})^\intercal\Delta_{i-1},
\end{aligned} \tag{4}$$

where **(a)** comes from a Taylor expansion. By the method of Lagrange multipliers and the observation that $\|\Delta\| = 1$ implies that $\|\Delta\|^2 = 1$, we have that any Δ_{i-1} that satisfies (4) must also satisfy:

$$\begin{aligned}
0 =& \frac{\partial}{\partial \Delta_{i-1}} \left(\frac{\partial f}{\partial \theta_{i-1}}(X_{i-1}, \theta_{i-1})^\intercal\Delta_{i-1} - \lambda\left(\|\Delta_{i-1}\|^2 - 1\right)\right) \\
=& \frac{\partial}{\partial \Delta_{i-1}} \left(\frac{\partial f}{\partial \theta_{i-1}}(X_{i-1}, \theta_{i-1})^\intercal\Delta_{i-1} - \lambda\left(\Delta_{i-1}^\intercal\Delta_{i-1} - 1\right)\right) \\
=& \frac{\partial f}{\partial \theta_{i-1}}(X_{i-1}, \theta_{i-1}) - 2\lambda\Delta_{i-1},
\end{aligned} \tag{5}$$

and so

$$\Delta_{i-1} = \frac{1}{2\lambda}\frac{\partial f}{\partial \theta_{i-1}}(X_{i-1}, \theta_{i-1}),$$

where λ is a scalar. It is straightforward to verify that this direction is the unique solution, Δ_{i-1}, and not just a critical point of the Lagrangian. If we ignore the scalar terms (e.g., by viewing them as part of the step size, α_i), we have the direction:

$$\Delta_{i-1} = \frac{\partial f}{\partial \theta_{i-1}}(X_{i-1}, \theta_{i-1}), \tag{6}$$

which is the direction of change to θ_{i-1} used by stochastic gradient descent in (3). It is also the direction of change to θ_{i-1} that is used by non-gradient learning rules, like temporal-difference learning (Sutton, 1988), that use updates of the form:

$$\theta_i = \theta_{i-1} + \delta_{i-1}\Delta_{i-1},$$

where here δ_{i-1} denotes an error term called the *temporal difference error*. In general, for learning rules that satisfy Assumption 1, the $\partial f(\cdot, \theta_{i-1}i)/\partial\theta_{i-1}$ terms denote different Δ_{i-1} terms, evaluated using different $x \in \mathcal{X}$.

The problem with learning rules that use (6)—learning rules that satisfy Assumption 1—is that they make an implicit assumption that the distance between $f(X_{i-1}, \theta_{i-1})$ and $f(X_{i-1}, \theta_{i-1}+\alpha\Delta\theta_i)$ should be measured using Euclidean distance in the parameters when selecting Δ_{i-1}. That is, they use $\|\Delta\|^2 \coloneqq \Delta^\intercal\Delta$ during the derivation of Δ_{i-1}—specifically to obtain (5) during the derivation.

The problem with using learning rules that satisfy Assumption 1, which use Euclidean distance in the parameters when deriving Δ_{i-1}, is that they intertwine the choices of which learning rule to use and which parameterized function to use. To see how this intertwining occurs, consider a parameterized function, g, that is congruent to f, with submersion ψ. Using f and Euclidean distance in the parameterization, the squared distance between $f(X_{i-1}, \theta_{i-1})$ and $f(X_{i-1}, \theta_{i-1} + \Delta))$ is $\Delta^\intercal\Delta$. However, using g and the Euclidean distance in the parameterization, the squared distance between the same two functions, $g(X_{i-1}, \psi(\theta_{i-1}))$ and $g(X_{i-1}, \psi(\theta_{i-1} + \Delta))$, is

$$\left(\psi(\theta_{i-1} + \Delta) - \psi(\theta_{i-1})\right)^\intercal \left(\psi(\theta_{i-1} + \Delta) - \psi(\theta_{i-1})\right),$$

which is not necessarily the same. These differing notions of distance will result in different solutions to (4), and thus different update directions. This is reflected by the fact that learning rules that satisfy Assumption 1 are not covariant or j-order covariant for any $j \in \mathbb{N}_{>0}$ and non-degenerate $\mathcal{G}$.

Furthermore, for some parameterizations, Euclidean distance in the parameters may be a poor notion of distance. For example, in a deep neural network, a weight at an early layer of the network may have little impact on the output of the network, while a weight near the output of the network might have a large impact. Using Euclidean distance in the parameters means that small changes to these two weights incur the same amount of distance, and so the direction of steepest ascent will favor larger changes to the weight later in the network, since small changes thereto can have a bigger influence on the network's output. Amari (1998) was the first to suggest that this line of reasoning could explain the tendency of algorithms for training neural networks to require many iterations of the learning rule to properly set the values of weights early in the network.

This raises the question: what notion of distance (or more generally, what dissimilarity function) should be used when computing Δ_{i-1}—the direction of steepest ascent of $f(X_{i-1}, \cdot)$ at θ_{i-1}? We would like to use (1), so that

$$\begin{aligned}\|\Delta_{i-1}\|^2 \coloneqq& \operatorname{dist}(\theta_{i-1}, \theta_{i-1} + \Delta_{i-1})^2 \\ =& \frac{1}{2}\int_{\mathcal{X}^2} (f(x, \theta_{i-1}) - f(x, \theta_{i-1} + \Delta_{i-1}))(f(y, \theta_{i-1}) - f(y, \theta_{i-1} + \Delta_{i-1})p(dx, dy),\end{aligned}$$

where $p(dx, dy) \coloneqq p_i(f(\cdot, \beta_i), \omega, z, dx, dy)$, where $z \in \mathcal{X}$ corresponds to z in (2). Although this definition of $\|\cdot\|$ is desirable, it does not ensure that a simple closed form exists for Δ_{i-1}. So, instead we use

$$\|\Delta_{i-1}\|^2 \coloneqq \tau_2\left(\operatorname{dist}(\theta_{i-1}, \theta_{i-1} + \cdot)^2, \theta_{i-1}, \theta_{i-1} + \Delta_{i-1}\right).$$

That is, we use a second order Taylor approximation of the dissimilarity function, d as our definition of squared distance. Although this second order Taylor approximation does *not* result in a definition of squared distance that yields covariant

updates, Theorem 1 shows that it is sufficient to yield first-order covariant updates. Also, notice that this use a second order Taylor approximation to a dissimilarity function is not unprecedented: Amari's natural gradient method using the Fisher information matrix equates to using a second order Taylor approximation of Kullback–Leibler divergence to measure squared distances when computing Δ_{i-1} (Thomas et al., 2016, Appendix A).

The use of a second-order Taylor approximation of $\mathrm{dist}(\theta_{i-1}, \theta_{i-1} + \cdot)^2$ results in a closed form for the Δ_{i-1} that satisfy (4) because:

$$\begin{aligned}\tau_2\left(\mathrm{dist}(\theta,\cdot)^2, \theta, \theta+\Delta\right) =& \underbrace{\mathrm{dist}(\theta,\theta)^2}_{=\text{(a)}} + \underbrace{\left(\frac{\partial\,\mathrm{dist}}{\partial\gamma}(\alpha,\gamma)^2\big|_{\substack{\alpha=\theta\\\gamma=\theta}}\right)^\intercal}_{=\text{(b)}}\Delta + \frac{1}{2}\Delta^\intercal\left(\frac{\partial^2\,\mathrm{dist}}{\partial\gamma^2}(\alpha,\gamma)^2\big|_{\substack{\alpha=\theta\\\gamma=\theta}}\right)\Delta \\ =& \frac{1}{2}\Delta^\intercal\left(\frac{\partial^2\,\mathrm{dist}}{\partial\gamma^2}(f,\alpha,\gamma)^2\big|_{\substack{\alpha=\theta\\\gamma=\theta}}\right)\Delta,\end{aligned}$$

since it is straightforward to verify that **(a)** and **(b)** are both zero.[2] Furthermore,

$$\left(\frac{\partial^2\,\mathrm{dist}}{\partial\gamma^2}(f,\alpha,\gamma)^2\big|_{\substack{\alpha=\theta\\\gamma=\theta}}\right) = \int_{\mathcal{X}^2}\frac{\partial f}{\partial\theta}(x,\theta)\frac{\partial f}{\partial\theta}(y,\theta)^\intercal p(dx,dy).$$

So,

$$\tau_2\left(\mathrm{dist}(\theta,\cdot)^2, \theta, \theta+\Delta\right) = \Delta^\intercal\left(\int_{\mathcal{X}^2}\frac{\partial f}{\partial\theta}(x,\theta)\frac{\partial f}{\partial\theta}(y,\theta)^\intercal p(dx,dy)\right)\Delta,$$

and thus

$$\|\Delta_{i-1}\|^2 \coloneqq \Delta_{i-1}^\intercal G_i^x(f,\theta_{i-1})\Delta_{i-1}.$$

Using this squared norm and the method of Lagrange multipliers as before, the solutions to (4) satisfy

$$\begin{aligned}0 =& \frac{\partial}{\partial\Delta_{i-1}}\left(\frac{\partial f}{\partial\theta_{i-1}}(x,\theta_{i-1})^\intercal\Delta_{i-1} - \frac{1}{2}\lambda\left(\Delta_{i-1}^\intercal G_i^x(f,\theta_{i-1})\Delta_{i-1} - 1\right)\right) \\ =& \frac{\partial f}{\partial\theta_{i-1}}(x,\theta_{i-1}) - \lambda G_i^x(f,\theta_{i-1})\Delta_{i-1},\end{aligned}$$

and so $\Delta_{i-1} = \frac{1}{\lambda}G_i^x(f,\theta_{i-1})^+\frac{\partial f}{\partial\theta_{i-1}}(x,\theta_{i-1})$, or ignoring the scalar terms as before (by viewing them as part of the step sizes),

$$\Delta_{i-1} = G_i^x(f,\theta_{i-1})^+\frac{\partial f}{\partial\theta_{i-1}}(x,\theta_{i-1}).$$

This definition of Δ_{i-1} is exactly what is used by $\tilde{l}$.

B. Proof of Theorem 1

We begin by establishing properties that we use later. Also, for brevity and to avoid clutter, we use several shorthand notations in all of the appendices: $\nabla\psi \coloneqq \frac{\partial\psi}{\partial\beta_i}(\beta_i)$, $\nabla f \coloneqq \frac{\partial f}{\partial\beta_i}(x,\beta_i)$, $\nabla^2 f \coloneqq \frac{\partial^2 f}{\partial\beta_i^2}(x,\beta_i)$, $\nabla g \coloneqq \frac{\partial g}{\partial\psi(\beta_i)}(x,\psi(\beta_i))$, and $\nabla^2 g \coloneqq \frac{\partial^2 g}{\partial\psi(\beta_i)^2}(x,\psi(\beta_i))$.

Property 1 (Jacobian Property)**.** *If f and g are congruent representations, then for all $x \in \mathcal{X}$ and $\theta \in \mathbb{R}^n$,*

$$\frac{\partial f}{\partial\theta}(x,\theta) = \left(\frac{\partial\psi}{\partial\theta}(\theta)\right)^\intercal\frac{\partial g}{\partial\psi(\theta)}(x,\psi(\theta)).$$

Proof.

$$\frac{\partial f}{\partial\theta}(x,\theta) \overset{\text{(a)}}{=} \frac{\partial g}{\partial\theta}(x,\psi(\theta)) = \left(\frac{\partial\psi}{\partial\theta}(\theta)\right)^\intercal\frac{\partial g}{\partial\psi(\theta)}(x,\psi(\theta)),$$

where **(a)** holds because $f(x,\theta) = g(x,\psi(\theta))$ for all $x \in \mathcal{X}$ and $\theta \in \mathbb{R}^n$ by the assumption that f and g are congruent representations. □

[2] Notice that here we have switched notation for differentiation. This is because $\frac{\partial d}{\partial\theta_{i-1}}(\theta_{i-1},\theta_{i-1})$ is ambiguous since the derivative is with respect to the second argument of d, not the first.

Property 2. *For all parameterized functions, $f \in \mathcal{P}$, all $g \in \mathcal{P}$ that are congruent to f, all $z \in \mathcal{X}$, all $\theta_0 \in \Theta^\iota$, all $\omega \in \Omega$, and all $i \in \mathbb{N}_{>0}$,*

$$G_i^z(f, \beta_i) = \nabla\psi^\intercal G_i^z(g, \psi(\beta_i))\nabla\psi.$$

Proof.

$$\begin{aligned}
G_i^z(f, \beta_i) &\coloneqq \int_{\mathcal{X}^2} \frac{\partial f(x, \beta_i)}{\partial \beta_i} \frac{\partial f(y, \beta_i)}{\partial \beta_i}^\intercal \, dp_i(f(\cdot, \beta_i), \omega, z, dx, dy) \\
&\overset{\text{(a)}}{=} \int_{\mathcal{X}^2} \nabla\psi^\intercal \frac{\partial g(x, \psi(\beta_i))}{\partial \psi(\beta_i)} \frac{\partial g(y, \psi(\beta_i))}{\partial \psi(\beta_i)}^\intercal \nabla\psi \, dp_i(f(\cdot, \beta_i), \omega, z, dx, dy) \\
&\overset{\text{(b)}}{=} \nabla\psi^\intercal \int_{\mathcal{X}^2} \frac{\partial g(x, \psi(\beta_i))}{\partial \psi(\beta_i)} \frac{\partial g(y, \psi(\beta_i))}{\partial \psi(\beta_i)}^\intercal \, dp_i(g(\cdot, \psi(\beta_i)), \omega, z, dx, dy)\nabla\psi \\
&= \nabla\psi^\intercal G_i^z(g, \psi(\beta_i))\nabla\psi,
\end{aligned}$$

where **(a)** comes from Property 1 and **(b)** holds because $\nabla\psi$ does not depend on x or y and because $f(\cdot, \beta_i) = g(\cdot, \psi(\beta_i))$ by the assumption that g is congruent to f. □

Property 3. *If l' is a first-order covariant update with respect to a sequence $(\beta_i)_{i=1}^\infty$, then for all $i \in \mathbb{N}_{>0}$, $\theta_0 \in \Theta^\iota$, and $\omega \in \Omega$,*

$$\nabla g^\intercal \nabla\psi \, (l_i'(f, \theta_0, \omega) - \beta_i) = \nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)).$$

Proof. We begin by writing out the Taylor expansions in the definition of first-order covariance:

$$\begin{aligned}
\tau_1(f(x, \cdot), \beta_i, l_i'(f, \theta_0, \omega)) &= \tau_1(g(x, \cdot), \psi(\beta_i), l_i'(g, \psi(\theta_0), \omega)) \\
f(x, \beta_i) + \nabla f^\intercal (l_i'(f, \theta_0, \omega) - \beta_i) &= g(x, \psi(\beta_i)) + \nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)) \\
\nabla f^\intercal (l_i'(f, \theta_0, \omega) - \beta_i) &\overset{\text{(a)}}{=} \nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)) \\
\nabla g^\intercal \nabla\psi \, (l_i'(f, \theta_0, \omega) - \beta_i) &\overset{\text{(b)}}{=} \nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)),
\end{aligned}$$

where **(a)** comes from the first terms on each side canceling by the definition of ψ and **(b)** comes from Property 1. □

To establish Theorem 1 we show that for all $g \in \mathcal{G}$ that are congruent to f,

$$\tau_1(f(x, \cdot), \beta_i, \tilde{l}_i(f, \theta_0, \omega)) = \tau_1(g(x, \cdot), \psi(\beta_i), \tilde{l}_i(g, \psi(\theta_0), \omega)). \tag{7}$$

To establish (7), we write out the Taylor expansions, as in the proof of Property 3. This gives an equality which, if satisfied, implies that $\tilde{l}$ is first-order covariant with respect to $(\beta_i)_{i=1}^\infty$ and $\mathcal{G}$.

$$\begin{aligned}
\tau_1(f(x, \cdot), \beta_i, \tilde{l}_i(f, \theta_0, \omega)) &= \tau_1(g(x, \cdot), \psi(\beta_i), \tilde{l}_i(g, \psi(\theta_0), \omega)) \\
f(x, \beta_i) + \nabla f^\intercal (\tilde{l}_i(f, \theta_0, \omega) - \beta_i) &= g(x, \psi(\beta_i)) + \nabla g^\intercal (\tilde{l}_i(g, \psi(\theta_0), \omega) - \psi(\beta_i)) \\
\nabla f^\intercal (\tilde{l}_i(f, \theta_0, \omega) - \beta_i) &\overset{\text{(a)}}{=} \nabla g^\intercal (\tilde{l}_i(g, \psi(\theta_0), \omega) - \psi(\beta_i)) \\
\nabla g^\intercal \nabla\psi \, (\tilde{l}_i(f, \theta_0, \omega) - \beta_i) &\overset{\text{(b)}}{=} \nabla g^\intercal (\tilde{l}_i(g, \psi(\theta_0), \omega) - \psi(\beta_i)).
\end{aligned} \tag{8}$$

We will show that this condition is met.

$$\begin{aligned}
&\nabla g^\intercal \nabla\psi(\tilde{l}_i(f, \theta_0, \omega) - \beta_i) \\
&\overset{\text{(a)}}{=} \nabla g^\intercal \nabla\psi \left(l_i'(f, \theta_0, \omega) - \beta_i + \int_{\mathcal{X}} G_i^\cdot(f, \beta_i)^+ \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i) \, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) \right) \\
&= \nabla g^\intercal \nabla\psi \, (l_i'(f, \theta_0, \omega) - \beta_i) + \nabla g^\intercal \nabla\psi \int_{\mathcal{X}} G_i^\cdot(f, \beta_i)^+ \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i) \, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) \\
&\overset{\text{(b)}}{=} \nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)) + \nabla g^\intercal \nabla\psi \int_{\mathcal{X}} G_i^\cdot(f, \beta_i)^+ \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i) \, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) \\
&\overset{\text{(c)}}{=} \nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)) + \nabla g^\intercal \nabla\psi \int_{\mathcal{X}} [\nabla\psi^\intercal G_i^\cdot(g, \psi(\beta_i))\nabla\psi]^+ \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i) \, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) j
\end{aligned} \tag{9}$$

where **(a)** comes from substituting in the definition of $\tilde{l}_i(f, \theta_{1:i}, \omega)$, **(b)** holds by Property 3, and **(c)** comes from Property 2. Notice that $\nabla\psi^\intercal \in \mathbb{R}^{n\times m}$ has full column rank (since $m \leq n$), $\nabla\psi \in \mathbb{R}^{m\times n}$ has full row rank, and $\text{rank}(G_i^x(g, \psi(\beta_i))) = m$ by the definition of $\mathcal{G}$ in Theorem 1. Thus, by Sylvester's rank inequality we have that $\text{rank}(G_i^x(g, \psi(\beta_i))\nabla\psi) \geq \text{rank}(G_i^x(g, \psi(\beta_i))) + \text{rank}(\nabla\psi) - m = m + m - m = m$. Also, due to its dimensions, $\text{rank}(G_i^x(g, \psi(\beta_i))\nabla\psi) \leq m$, and so we can conclude that $\text{rank}(G_i^x(g, \psi(\beta_i))\nabla\psi) = m$. So, $\nabla\psi^\intercal$ has full column rank and $G_i^x(g, \psi(\beta_i))\nabla\psi$ has full row rank. So, by two applications of the rule that $(AB)^+ = B^+A^+$ if A has full column rank and B has full row rank (Greville & Nall, 1966), we have that:

$$(\nabla\psi^\intercal G_i^\cdot(g, \psi(\beta_i))\nabla\psi)^+ = \nabla\psi^+ G_i^\cdot(g, \psi(\beta_i))^+ (\nabla\psi^\intercal)^+.$$

Continuing (9), we therefore have that:

$$\begin{aligned}
&\nabla g^\intercal \nabla\psi(\tilde{l}_i(f, \theta_0, \omega) - \beta_i) \\
=&\nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)) + \nabla g^\intercal \nabla\psi \int_{\mathcal{X}} \nabla\psi^+ G_i^\cdot(g, \psi(\beta_i))^+ (\nabla\psi^\intercal)^+ \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) \\
=&\nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)) + \nabla g^\intercal \nabla\psi\nabla\psi^+ \int_{\mathcal{X}} G_i^\cdot(g, \psi(\beta_i))^+ (\nabla\psi^\intercal)^+ \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) \\
\overset{\text{(a1)}}{=}&\nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)) + \nabla g^\intercal \int_{\mathcal{X}} G_i^\cdot(g, \psi(\beta_i))^+ (\nabla\psi^\intercal)^+ \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) \\
\overset{\text{(b)}}{=}&\nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)) + \nabla g^\intercal \int_{\mathcal{X}} G_i^\cdot(g, \psi(\beta_i))^+ (\nabla\psi^\intercal)^+ \nabla\psi^\intercal \frac{\partial g}{\partial \psi(\beta_i)}(\cdot, \psi(\beta_i))\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) \\
\overset{\text{(a2)}}{=}&\nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)) + \nabla g^\intercal \int_{\mathcal{X}} G_i^\cdot(g, \psi(\beta_i))^+ \frac{\partial g}{\partial \psi(\beta_i)}(\cdot, \psi(\beta_i))\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) \\
\overset{\text{(c)}}{=}&\nabla g^\intercal (l_i'(g, \psi(\theta_0), \omega) - \psi(\beta_i)) + \nabla g^\intercal \int_{\mathcal{X}} G_i^\cdot(g, \psi(\beta_i))^+ \frac{\partial g}{\partial \psi(\beta_i)}(\cdot, \psi(\beta_i))\, d\mu_i(g, \psi(\theta_0), \omega, \cdot) \\
=&\nabla g^\intercal \left(\tilde{l}_i(g, \psi(\theta_0), \omega) - \psi(\beta_i)\right),
\end{aligned} \tag{10}$$

where **(a1)** and **(a2)** hold because $\nabla\psi$ has linearly independent rows because it is full rank, and has more columns than rows by the requirement that $m \leq n$ in the definition of congruent representations, and so $\nabla\psi^+$ is a right-inverse, **(b)** holds by Property 1 and **(c)** holds because $f(\cdot, \beta_i) = g(\cdot, \psi(\beta_i))$ by the assumption that g is congruent to f. Notice that (10) is equal to the right side of (8), and so we conclude.

C. Proof of Theorem 2

Since $w^\star$ is a critical point:

$$\begin{aligned}
0 &= \int_{\mathcal{X}} (1 - \hat{1}(\cdot, w^\star)) \frac{\partial \hat{1}}{\partial w^\star}(\cdot, w^\star)\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) \\
0 &= \int_{\mathcal{X}} \left(1 - (w^\star)^\intercal \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)\right) \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot).
\end{aligned}$$

Rearranging terms, we obtain a new expression that is equal to a term in the learning rule, l:

$$\int_{\mathcal{X}} \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) = \int_{\mathcal{X}} \left((w^\star)^\intercal \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)\right) \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) \tag{11}$$

$$= \int_{\mathcal{X}} \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i) \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)^\intercal\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot) w^\star, \tag{12}$$

Replacing the left side of (11) in a learning rule, l, that satisfies Assumption 1, with (12), we have that l can be written as:

$$l_i(f, \theta_0, \omega) = l_i'(f, \theta_0, \omega) + \left[\int_{\mathcal{X}} \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i) \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)^\intercal\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot)\right] w^\star.$$

Similarly, $\tilde{l}$ from Theorem 1 can be written as

$$\tilde{l}_i(f, \theta_0, \omega) = l_i'(f, \theta_0, \omega) + \left[\int_{\mathcal{X}} G_i^\cdot(f, \beta_i)^+ \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i) \frac{\partial f}{\partial \beta_i}(\cdot, \beta_i)^\intercal\, d\mu_i(f(\cdot, \beta_i), \omega, \cdot)\right] w^\star. \tag{13}$$

Since

$$G_i^x(f,\beta_i) = \int_{\mathcal{X}} \frac{\partial f}{\partial \beta_i}(\cdot,\beta_i)\frac{\partial f}{\partial \beta_i}(\cdot,\beta_i)^\intercal \, d\mu_i(f(\cdot,\beta_i),\omega,\cdot),$$

or

$$G_i^x(f,\beta_i) = \frac{\partial f}{\partial \beta_i}(x,\beta_i)\frac{\partial f}{\partial \beta_i}(x,\beta_i)^\intercal,$$

and $G_i^x(f,\beta_i)$ is full rank, terms in (13) cancel to give:

$$\tilde{l}_i(f,\theta_0,\omega) = l_i'(f,\theta_0,\omega) + w^\star.$$

D. Proof of Theorem 3

We show that every learning rule, l, that is second-order covariant with respect to any sequence, $(\beta_i)_{i=1}^\infty$, and a set $\mathcal{G}$, must use the trivial update, $l_i(f,\theta_0,\omega) \coloneqq \beta_i$ for all parameterized functions, f, where **1)** $n = k = 1$, **2)** both $g(x,\theta) \coloneqq f(x,\ln(\theta))$ and $h(x,\theta) \coloneqq f(x,\ln(\theta)/2)$ are in $\mathcal{G}$ and congruent to f and **3)** both $\frac{\partial g}{\partial \theta}(\cdot,\beta_i)$ and $\frac{\partial^2 g}{\partial \theta^2}(\cdot,\beta_i)$ are not collinear and $\frac{\partial h}{\partial \theta}(\cdot,\beta_i)$ and $\frac{\partial^2 h}{\partial \theta^2}(\cdot,\beta_i)$ are not collinear.

To show this result, we will assume that l is a second-order covariant learning rule and will then show that, under these conditions, $l_i(f,\theta_0,\omega) \coloneqq \beta_i$. Since l is second-order covariant with respect to $(\beta_i)_{i=1}^\infty$ and $\mathcal{G}$, we have that:

$$\tau_2(f(x,\cdot),\beta_i,l_i(f,\theta_0,\omega)) = \tau_2(g(x,\cdot),\psi(\beta_i),l_i(g,\psi(\theta_0),\omega)) = \tau_2(h(x,\cdot),\phi(\beta_i),l_i(h,\phi(\theta_0),\omega)),$$

and so:

$$a\nabla f + \frac{a^2}{2}\nabla^2 f = b\nabla g + \frac{b^2}{2}\nabla^2 g = c\nabla h + \frac{c^2}{2}\nabla^2 h, \tag{14}$$

where $a \coloneqq l_i(f,\theta_0,\omega) - \beta_i$, $b \coloneqq l_i(g,\psi(\theta_0),\omega) - \psi(\beta_i)$, $c \coloneqq l_i(h,\phi(\theta_0),\omega) - \phi(\beta_i)$, $\nabla h \coloneqq \frac{\partial h}{\partial \phi(\beta_i)}(x,\phi(\beta_i))$, and $\nabla^2 h \coloneqq \frac{\partial^2 h}{\partial \phi(\beta_i)^2}(x,\phi(\beta_i))$.

We will show that, given f and the g and h specified in the theorem, (14) is only satisfied by $a = 0$, $b = 0$, and $c = 0$, which by the definitions of a, b, and c implies our result. Specifically, let:

$$g(\psi(\theta)) \coloneqq f(\ln(\psi(\theta)))$$
$$h(\phi(\theta)) \coloneqq f\left(\frac{1}{2}\ln(\phi(\theta))\right).$$

So, g and h are congruent to f with submersions $\psi(\theta) = e^\theta$ and $\phi(\theta) = e^{2\theta}$, respectively. Thus, we have the following:

$$\nabla\psi = e^{\beta_i}$$
$$\nabla\phi = 2e^{2\beta_i}$$
$$\nabla^2\psi = e^{\beta_i}$$
$$\nabla^2\phi = 4e^{2\beta_i}$$
$$\nabla f = \nabla\psi\nabla g = e^{\beta_i}\nabla g$$
$$\nabla f = \nabla\phi\nabla h = 2e^{2\beta_i}\nabla h$$
$$\nabla^2 f = \nabla^2 g\nabla\psi^2 + \nabla g\nabla^2\psi = e^{2\beta_i}\nabla^2 g + e^{\beta_i}\nabla g$$
$$\nabla^2 f = \nabla^2 h\nabla\phi^2 + \nabla h\nabla^2\phi = 4e^{4\beta_i}\nabla^2 h + 4e^{2\beta_i}\nabla h.$$

From (14) we have the requirement that for all $x \in \mathcal{X}$:

$$\begin{aligned} b\nabla g + \frac{b^2}{2}\nabla^2 g =& a\nabla f + \frac{a^2}{2}\nabla^2 f \\ =& ae^{\beta_i}\nabla g + \frac{a^2}{2}\left(e^{2\beta_i}\nabla^2 g + e^{\beta_i}\nabla g\right) \\ =& \left(ae^{\beta_i} + \frac{a^2}{2}e^{\beta_i}\right)\nabla g + \frac{a^2}{2}e^{2\beta_i}\nabla^2 g. \end{aligned} \tag{15}$$

Recall that ∇g and $\nabla^2 g$ (and ∇h and $\nabla^2 h$) are not collinear functions. Thus, the only way for (15) to hold for all x is if

$$b = ae^{\beta_i} + \frac{a^2}{2}e^{\beta_i}, \tag{16}$$

and

$$\frac{b^2}{2} = \frac{a^2}{2}e^{2\beta_i}. \tag{17}$$

Similarly, from (14) we have the requirement that for all $x \in \mathcal{X}$:

$$\begin{aligned} c\nabla h + \frac{c^2}{2}\nabla^2 h =& a\nabla f + \frac{a^2}{2}\nabla^2 f \\ =& 2ae^{2\beta_i}\nabla h + \frac{a^2}{2}\left(4e^{4\beta_i}\nabla^2 h + 4e^{2\beta_i}\nabla h\right) \\ =& \left(2ae^{2\beta_i} + \frac{a^2}{2}4e^{2\beta_i}\right)\nabla h + a^2 2e^{4\beta_i}\nabla^2 h, \end{aligned}$$

and thus we have that

$$c = 2ae^{2\beta_i} + \frac{a^2}{2}4e^{2\beta_i}, \tag{18}$$

and

$$\frac{c^2}{2} = a^2 2e^{4\beta_i}. \tag{19}$$

It is straightforward to verify using a computer algebra system like Wolfram Alpha that the only values for a, b, c that satisfy (16), (17), (18), and (19) simultaneously occur when $a = b = c = 0$. Since $a = b = c = 0$ corresponds to the trivial learning rule, $l_i(f, \theta_0, \omega) = \beta_i$, we conclude.