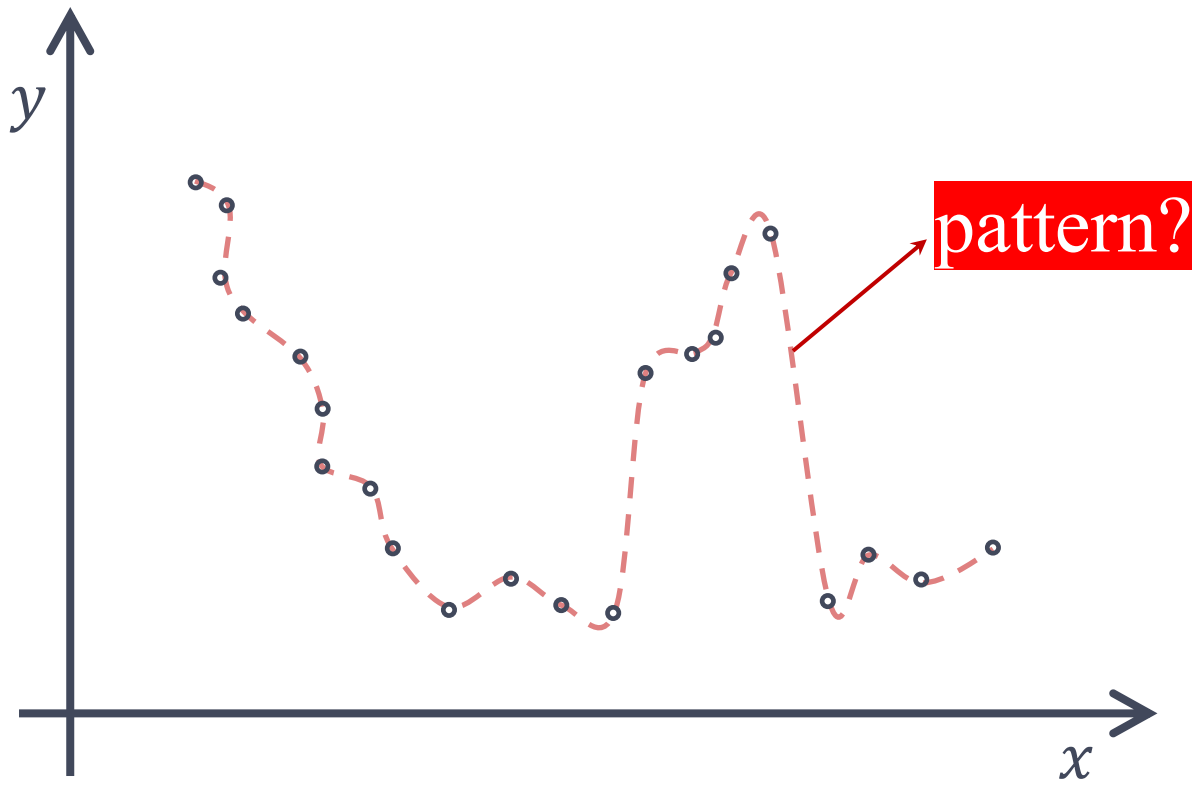
Covariate-Guided Clusterwise Linear Regression for Generalization to Unseen Data

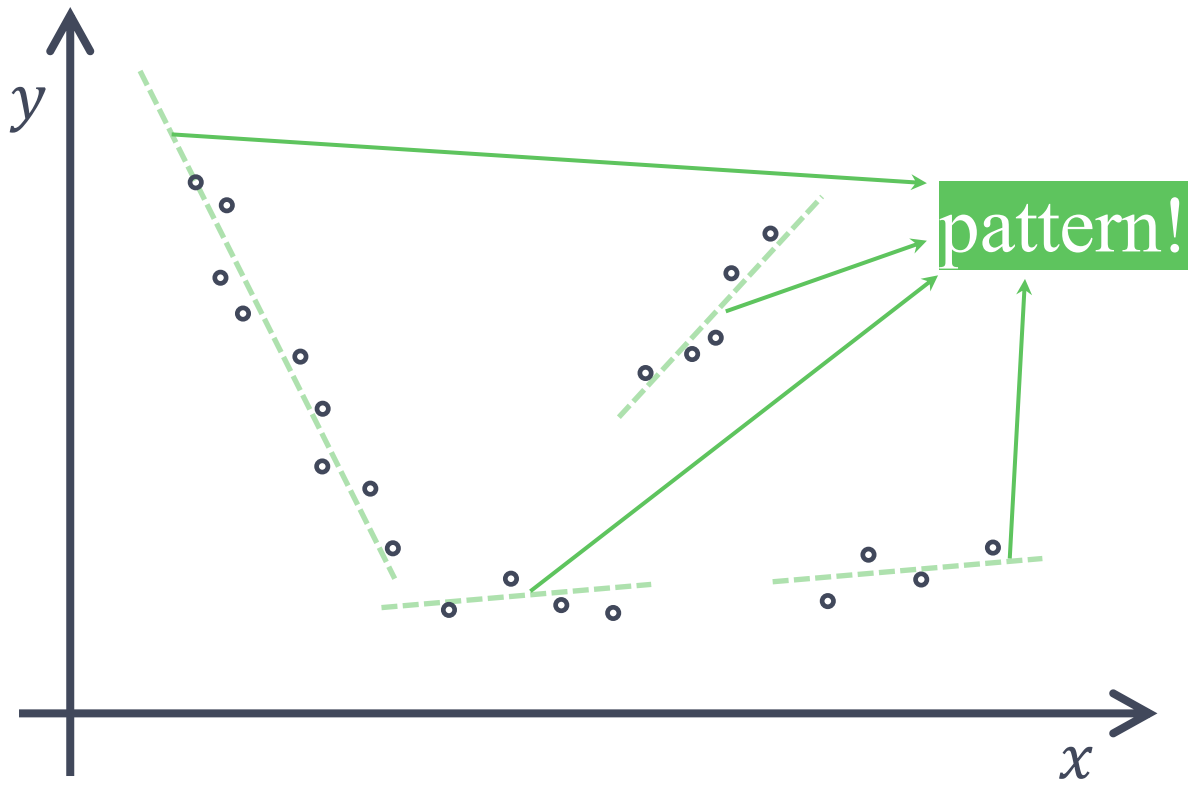
Dohyun Bu Hyunho Kim Jong-Seok Lee

Why Local Linear Models?



Many real-world data exhibit **globally nonlinear**, but **locally simple** patterns.

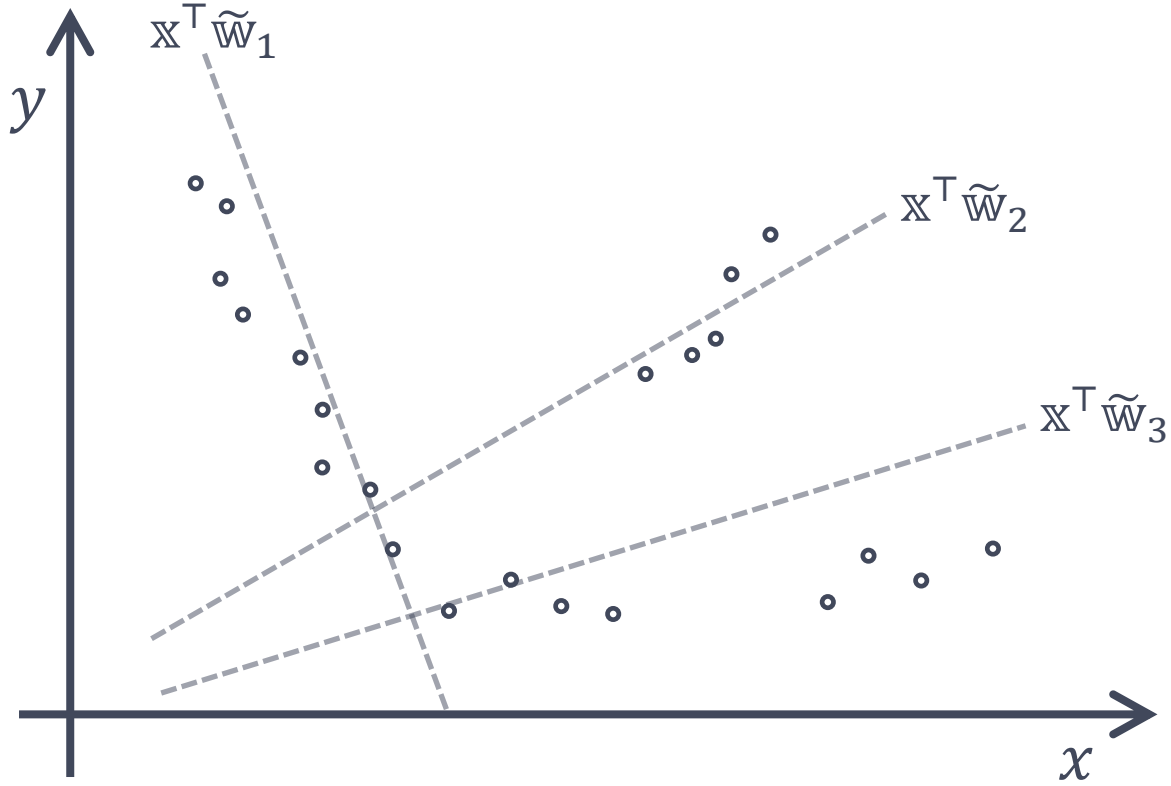
Why Local Linear Models?



Many real-world data exhibit **globally nonlinear**, but **locally simple** patterns.

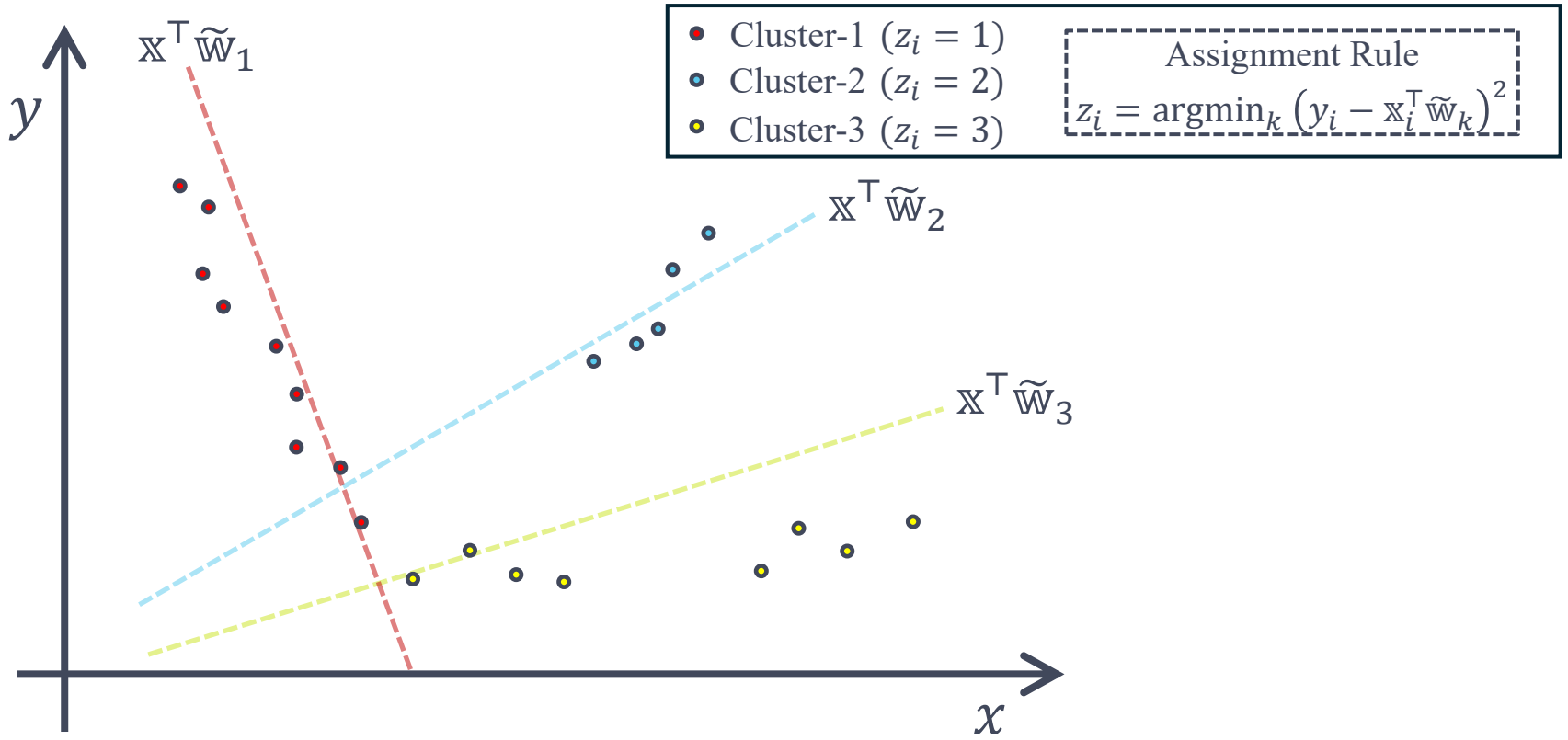
➔ This motivates Clusterwise Linear Regression (**CLR**)

CLR Learns Local Experts,



{ **C**lusterwise **L**inear **R**egression (**CLR**) }

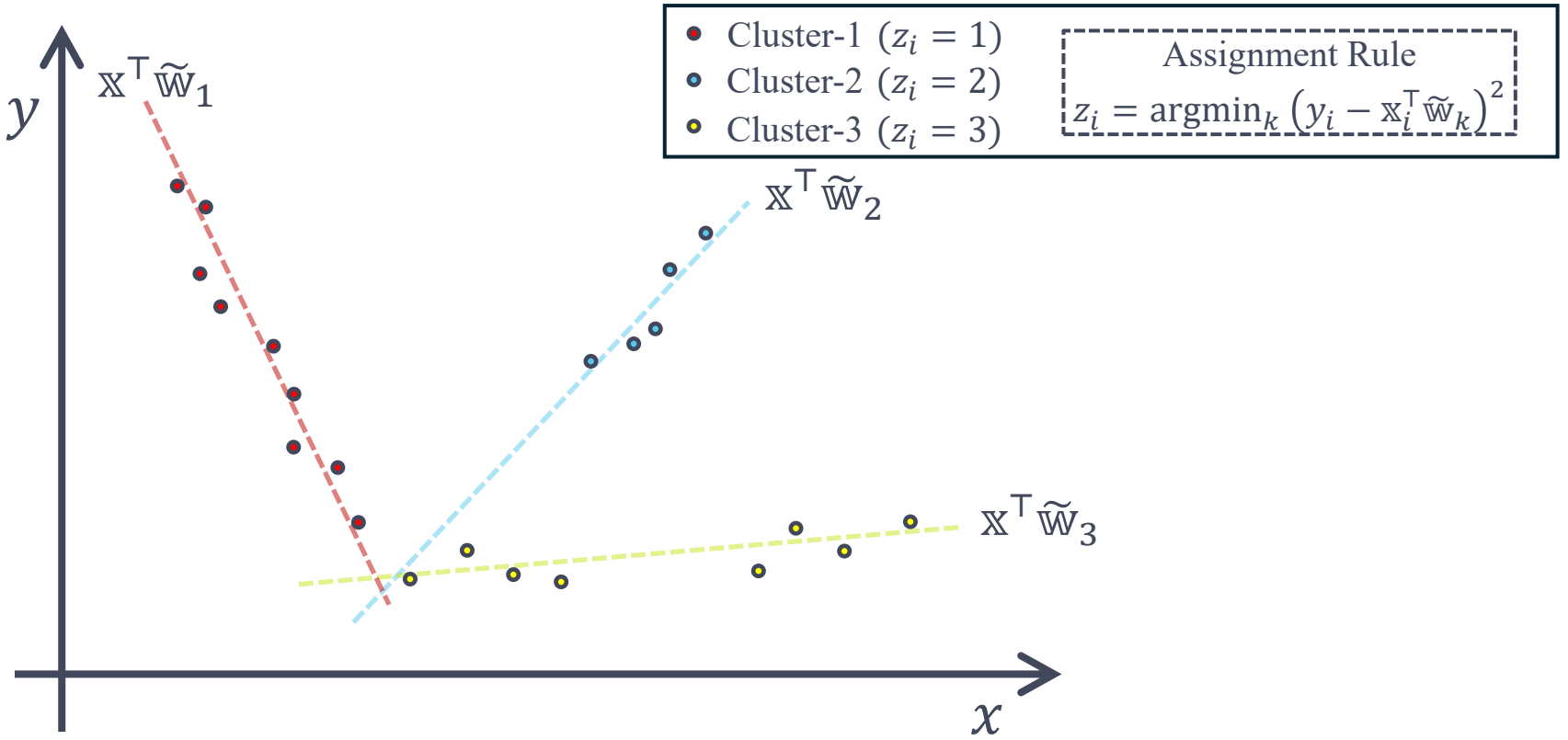
CLR Learns Local Experts,



{ Clusterwise Linear Regression (CLR) }

Step 1. Response-aware assignment

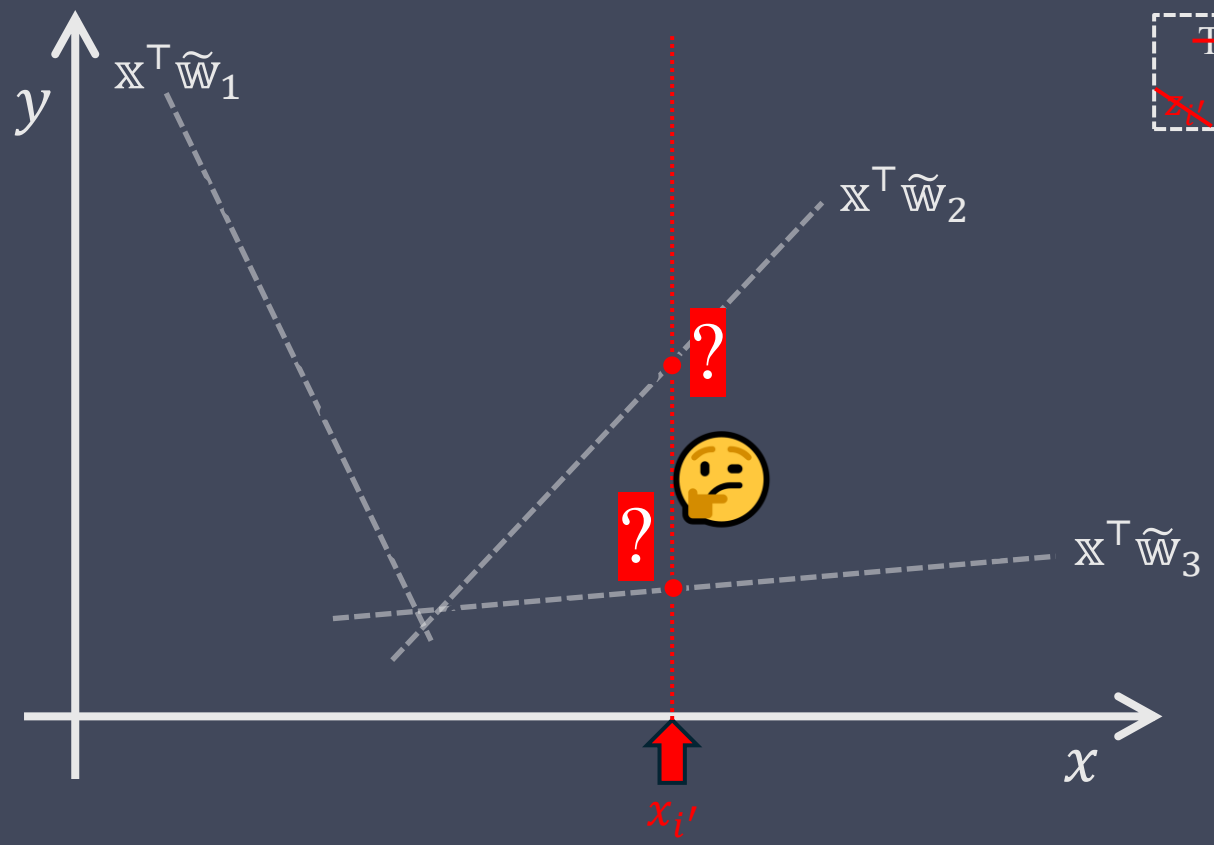
CLR Learns Local Experts,



{ **C**lusterwise **L**inear **R**egression (**CLR**) }

Step 2. Update each expert using assigned points Stable convergence! 😀

CLR Learns Local Experts, But Not Test-time Routing



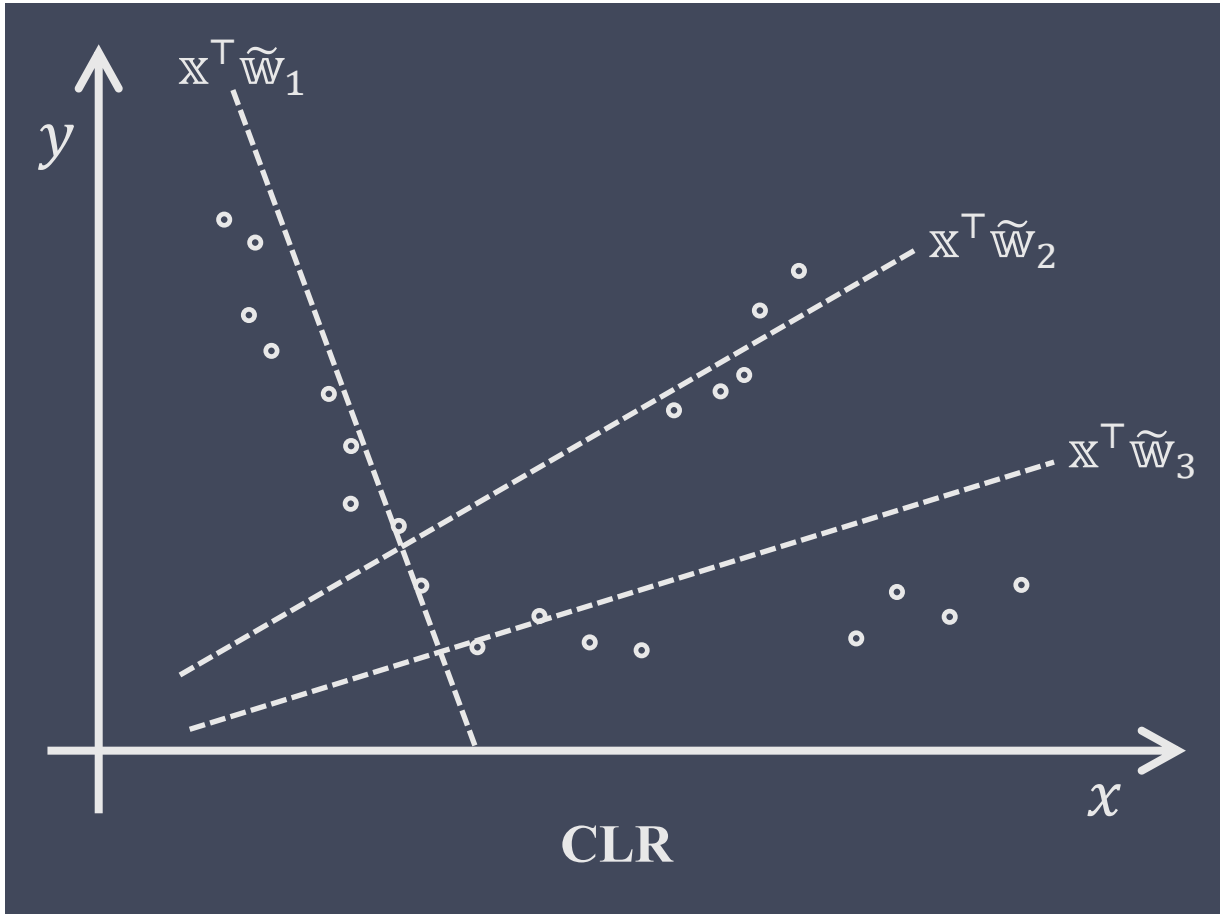
~~Test-time Assignment Rule~~

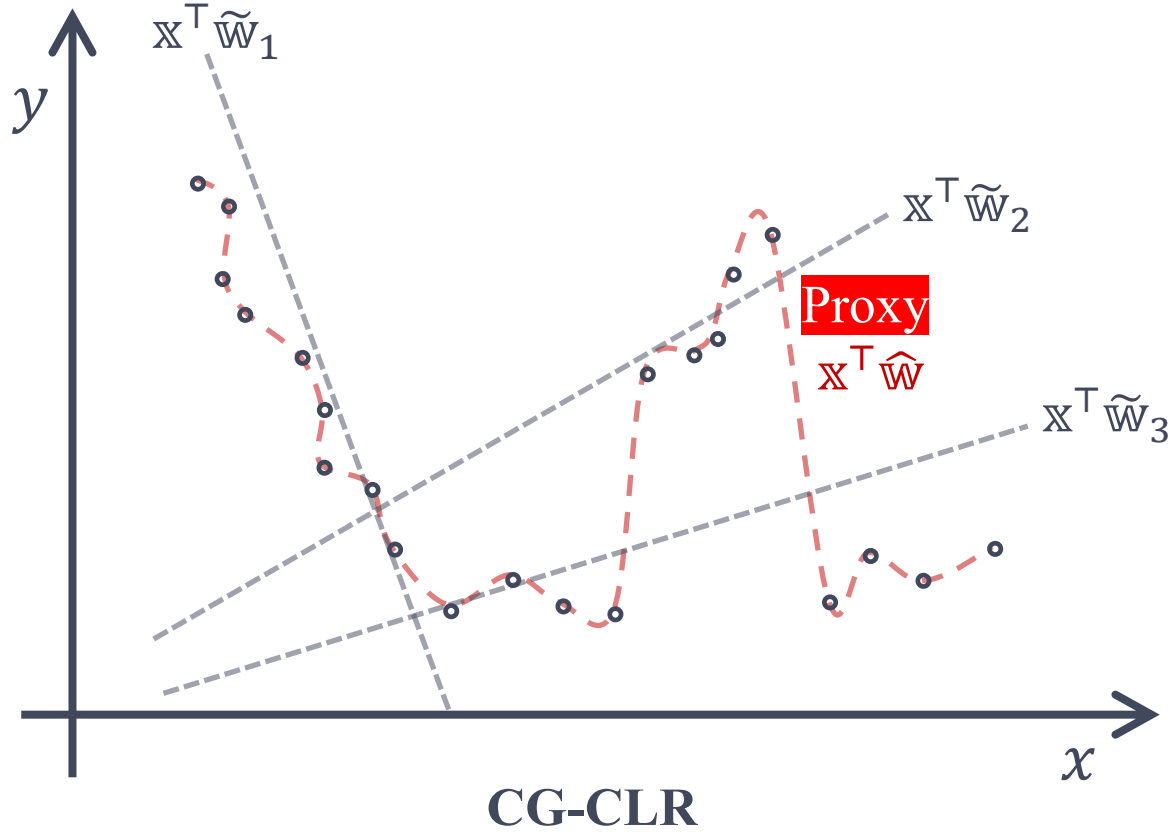
~~$z_{i'}$~~ $= \text{argmin}_k \left(~~$y_{i'}$~~ - \mathbb{x}_{i'}^\top \widetilde{\mathbb{w}}_k \right)^2$

For a new covariate $x_{i'}$ and unknown $y_{i'}$, we cannot assign the cluster-$z_{i'}$.

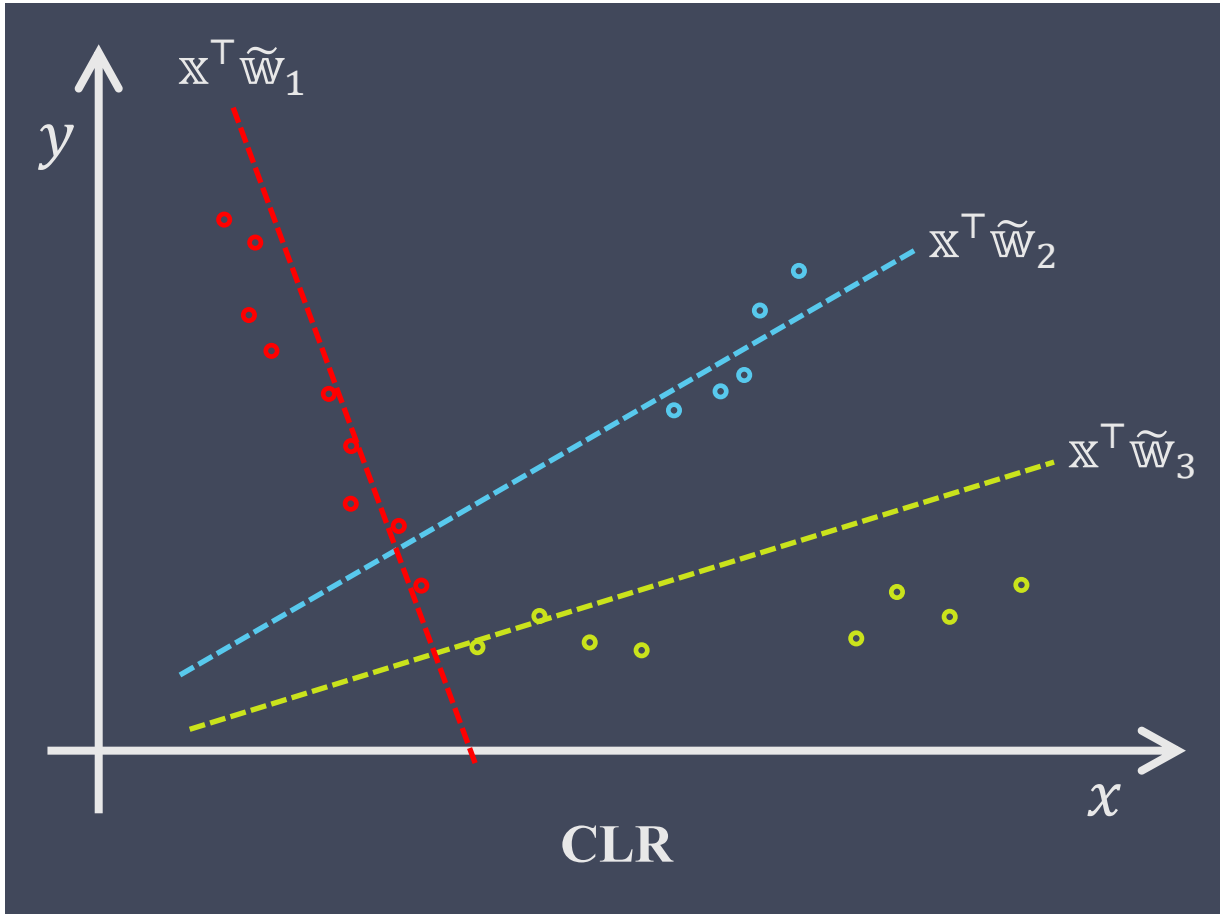
"Response-aware training, **infeasible** test-time prediction"

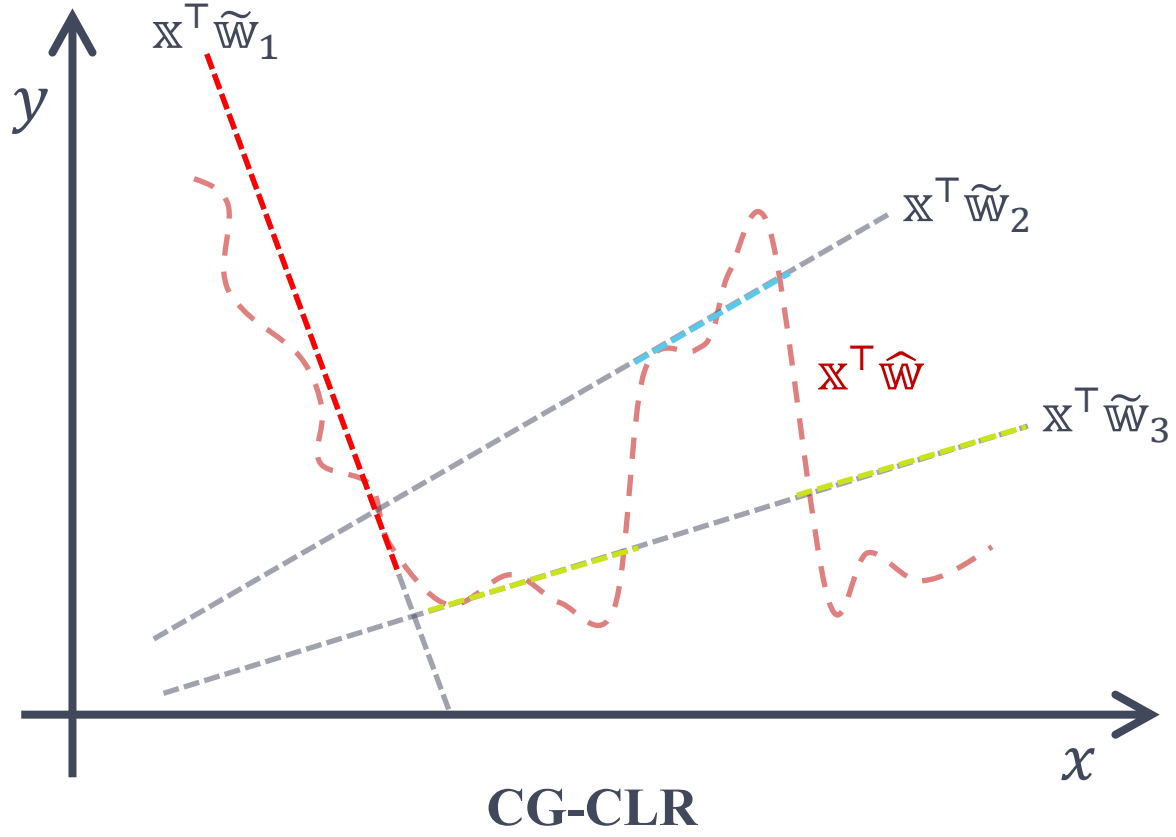
CG-CLR Adds Test-time Routing via a Proxy!





CG-CLR Adds Test-time Routing via a Proxy!





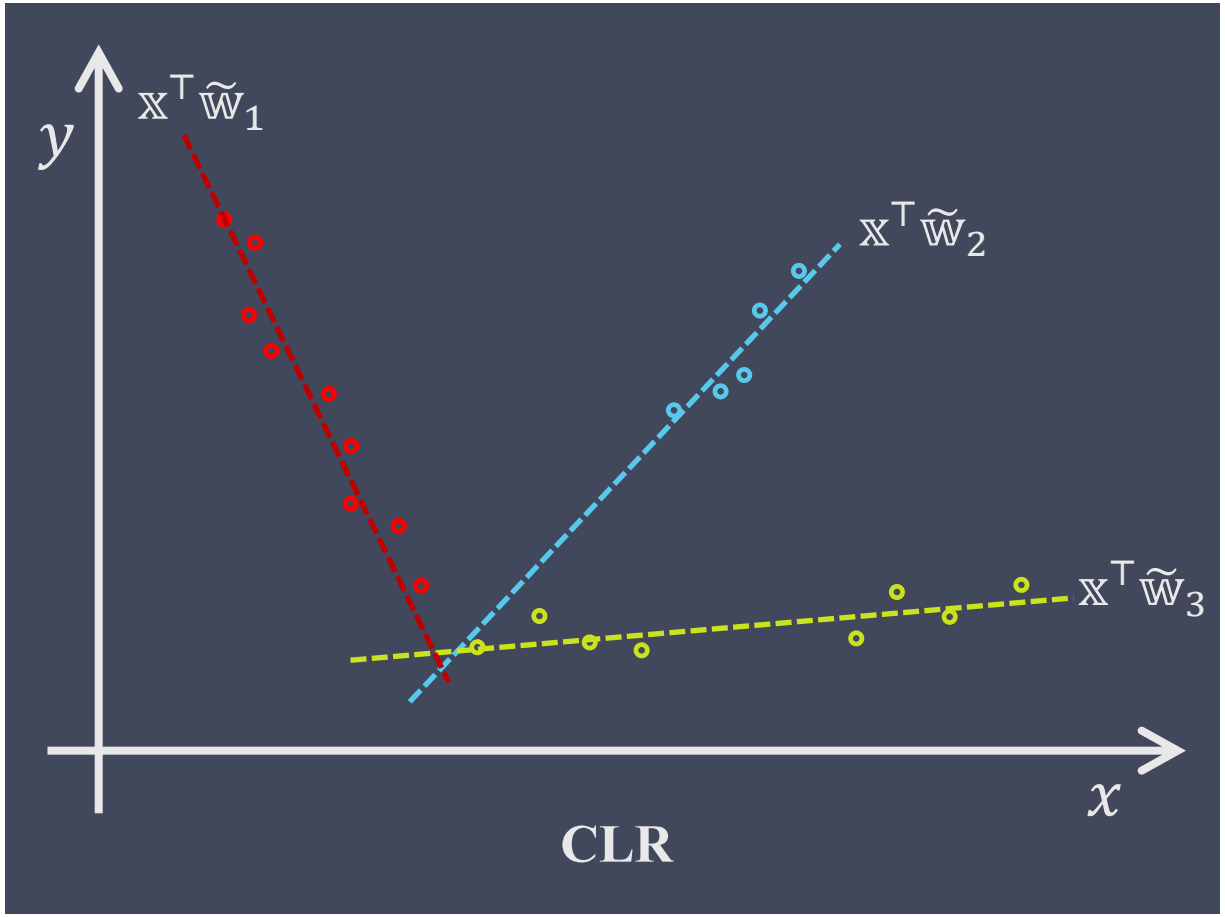
Response-aware Assignment

$$z_i = \text{argmin}_k \left(y_i - \mathbb{x}_i^\top \widetilde{\mathbb{w}}_k\right)^2$$

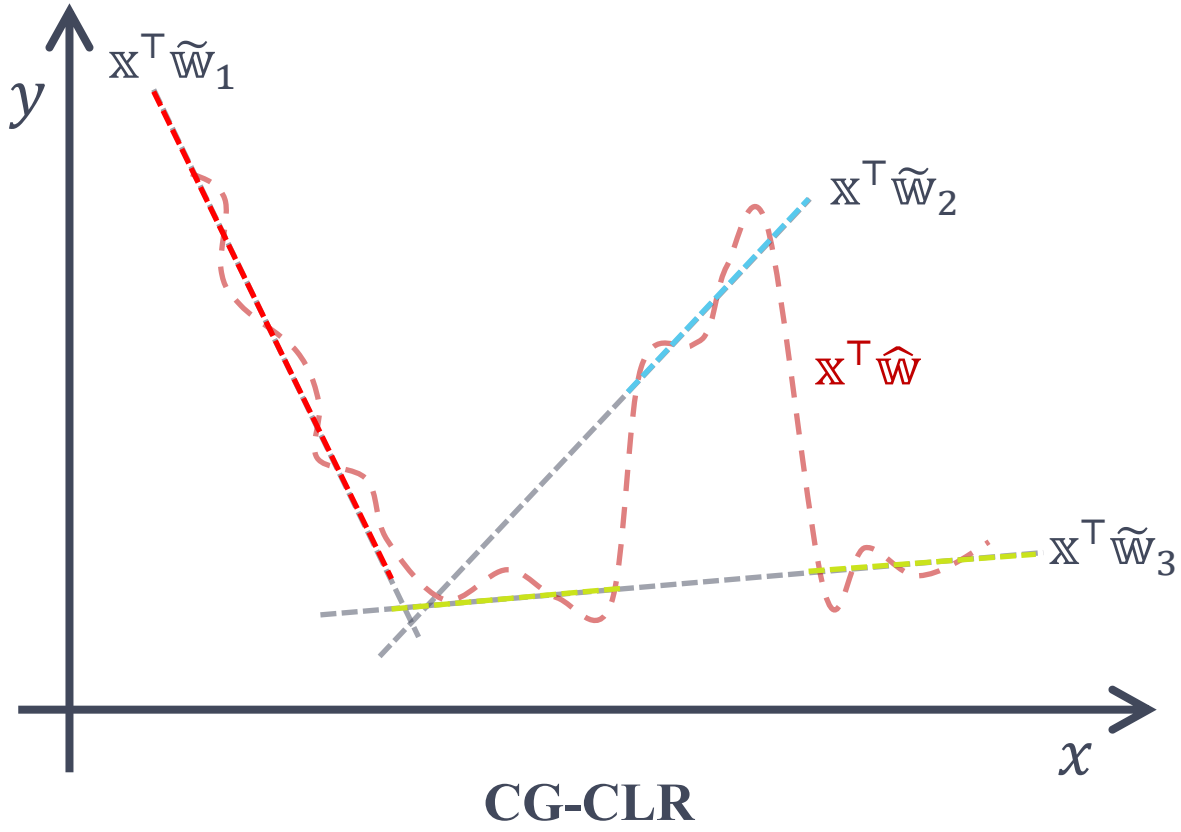
Proxy-based Assignment

$$z_i = \text{argmin}_k \left(\mathbb{x}_i^\top \widehat{\mathbb{w}}_i - \mathbb{x}_i^\top \widetilde{\mathbb{w}}_k\right)^2$$

CG-CLR Adds Test-time Routing via a Proxy!



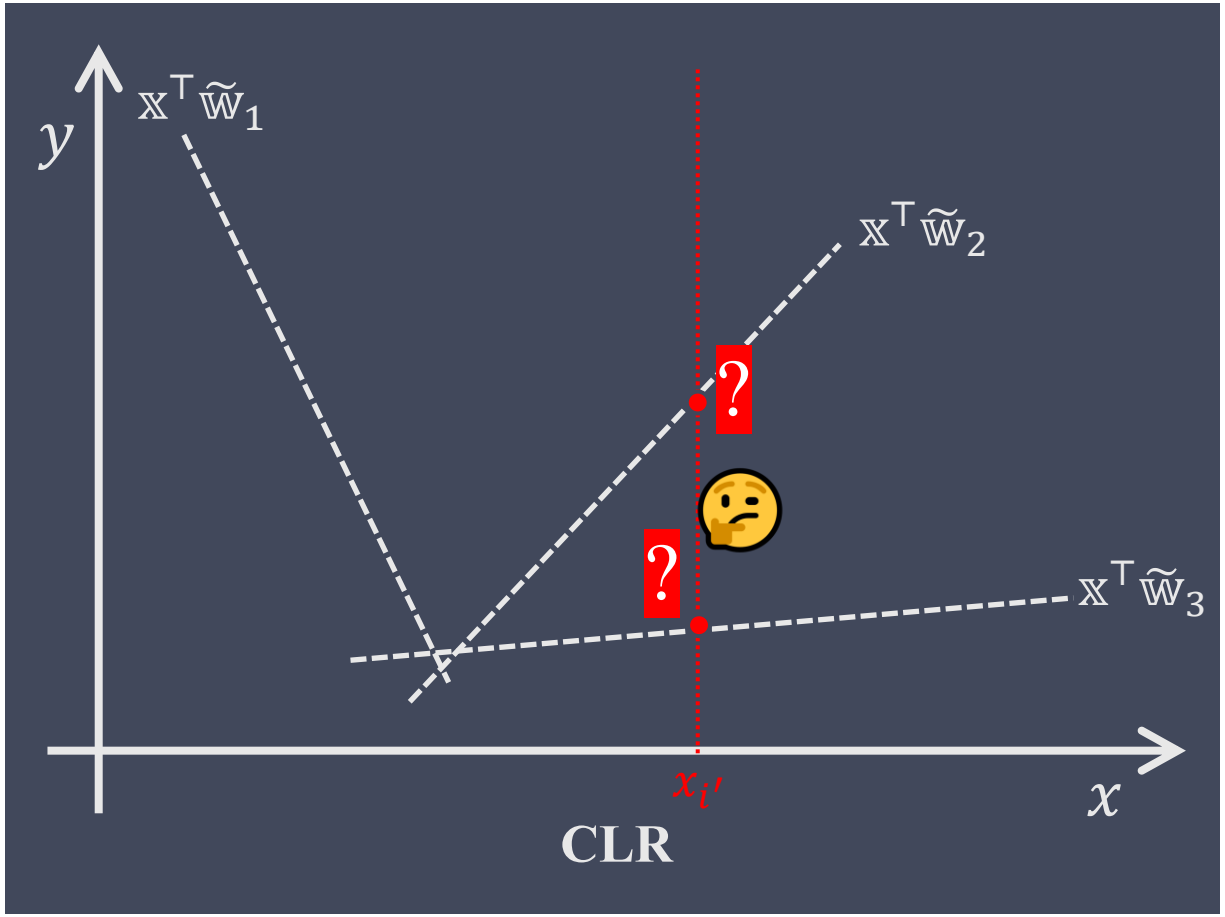
Update each expert using assigned points

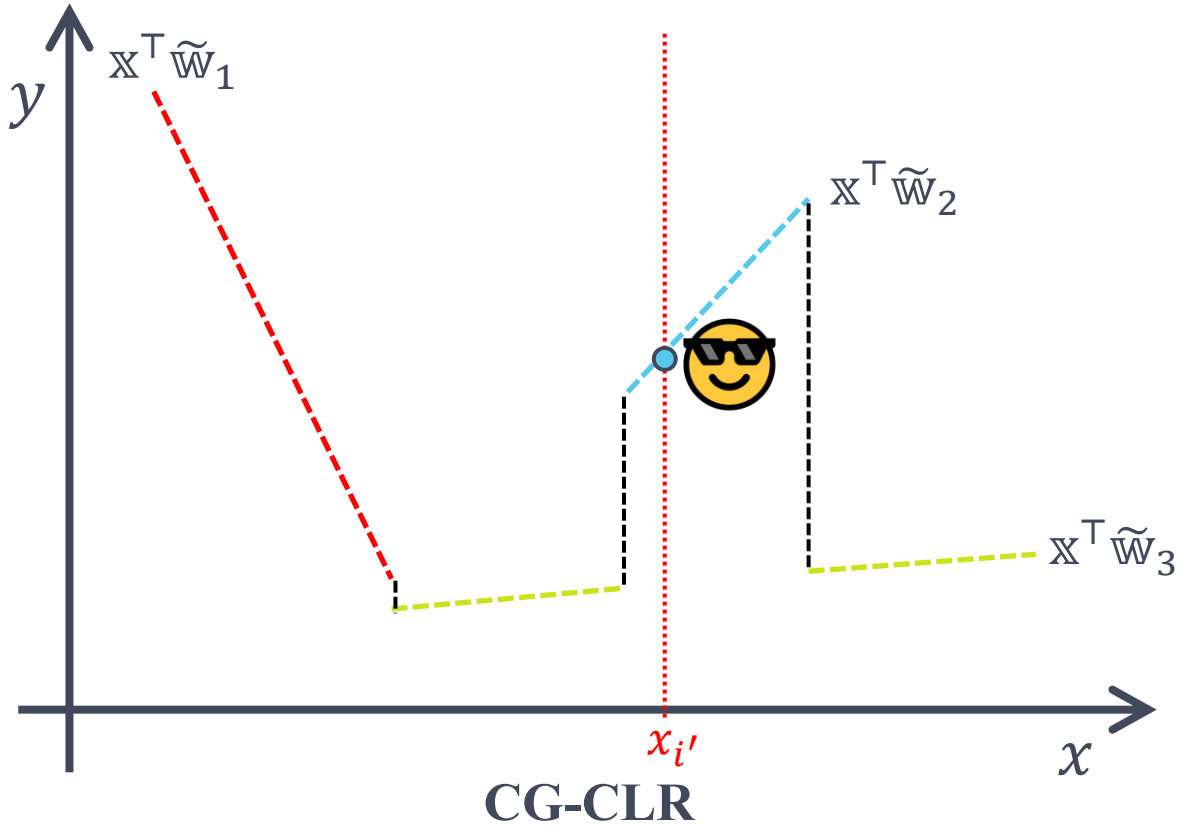


Update each expert using assigned **proxy outputs**

Stable convergence! 😀

CG-CLR Adds Test-time Routing via a Proxy!

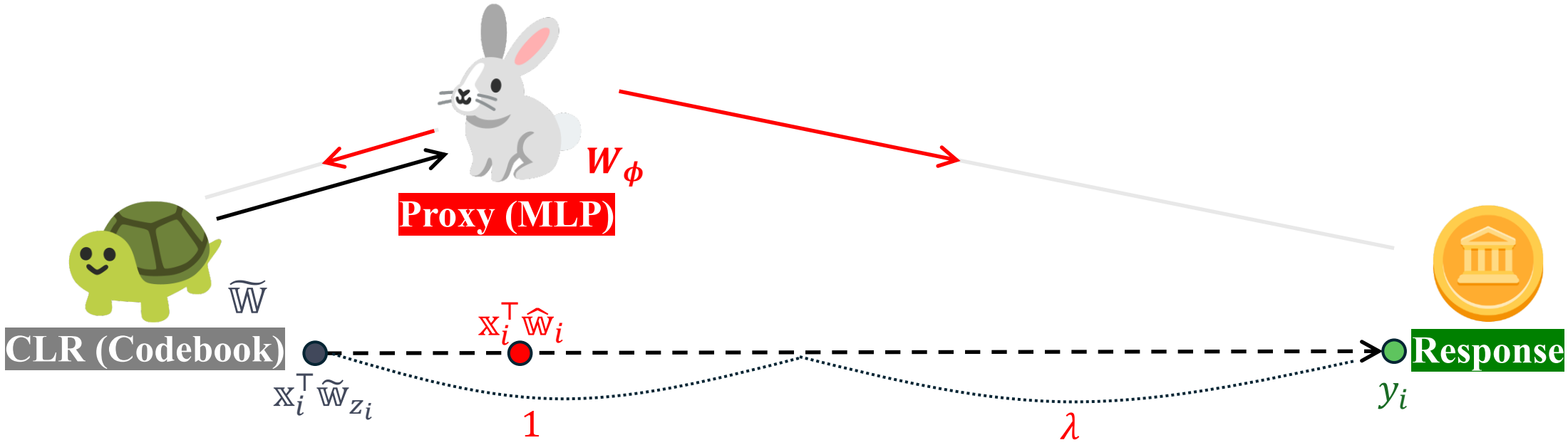


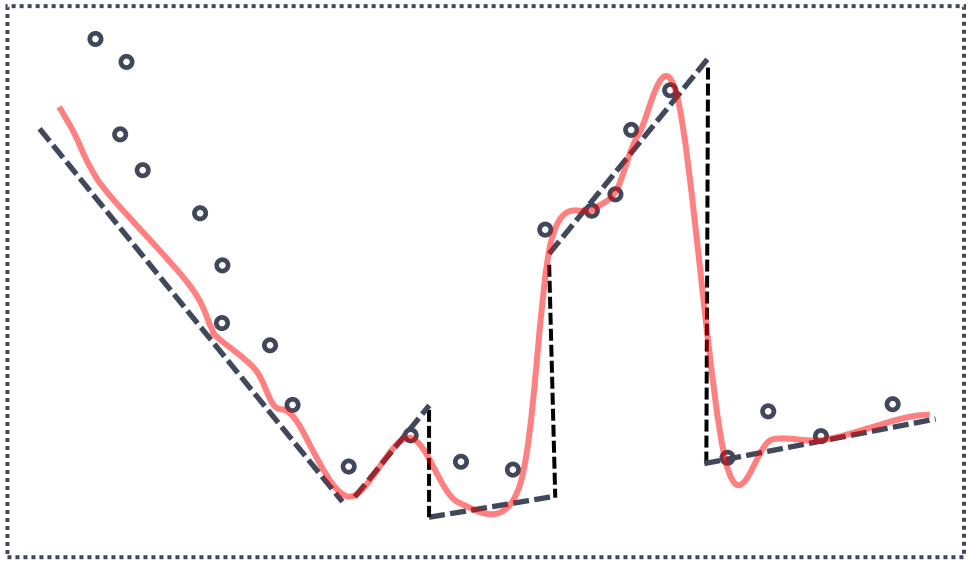


Covariate-guided test-time prediction!

without response $y_{i'}$

Why does the Optimization stay Stable?

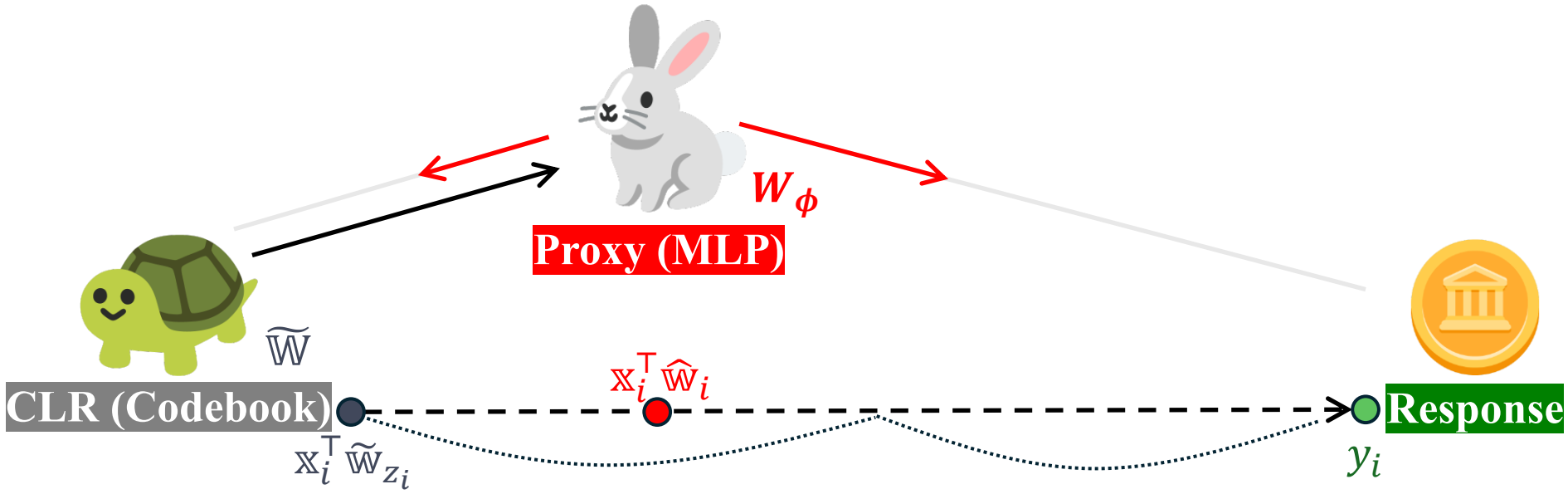


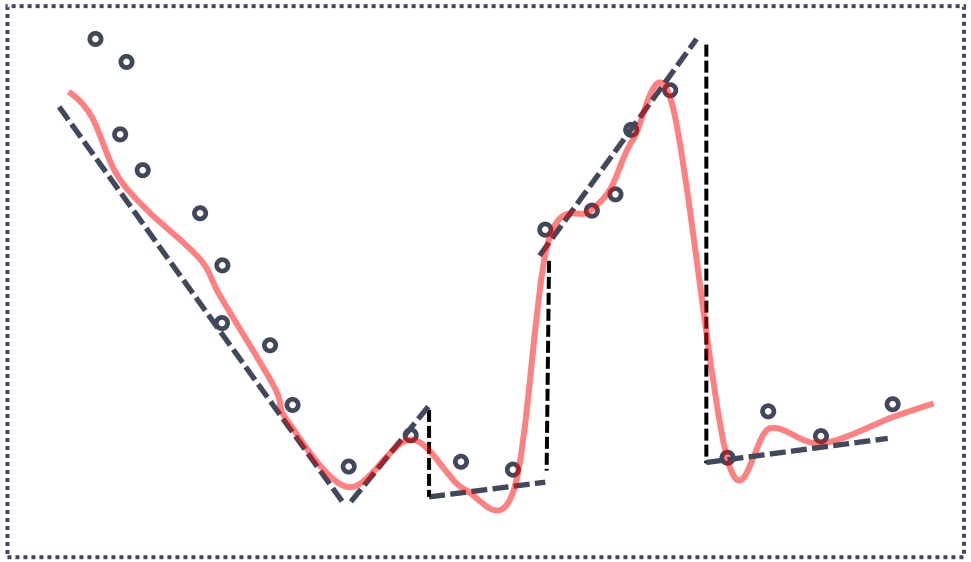


$$\phi^{(t+1)} \leftarrow \phi^{(t)} + \frac{2\eta(1+\lambda)}{N} \sum_{i=1}^{N} \left(\frac{y_i + \lambda \mathbb{x}_i^\top \tilde{\mathbb{w}}_{z_i}^{(t)}}{1+\lambda} - \mathbb{x}_i^\top \hat{\mathbb{w}}_i^{(t)} \right) (\nabla_\phi \hat{\mathbb{w}}_i^{(t)})^\top \mathbb{x}_i,$$

$$\tilde{\mathbb{w}}_j^{(t+1)} \leftarrow \tilde{\mathbb{w}}_j^{(t)} + \frac{2\eta(1+\lambda)}{N} \sum_{i \in S_j^{(t)}} \mathbb{x}_i \mathbb{x}_i^\top \left(\hat{\mathbb{w}}_i^{(t)} - \tilde{\mathbb{w}}_j^{(t)} \right), \quad \forall j \in [K].$$

Why does the Optimization stay Stable?

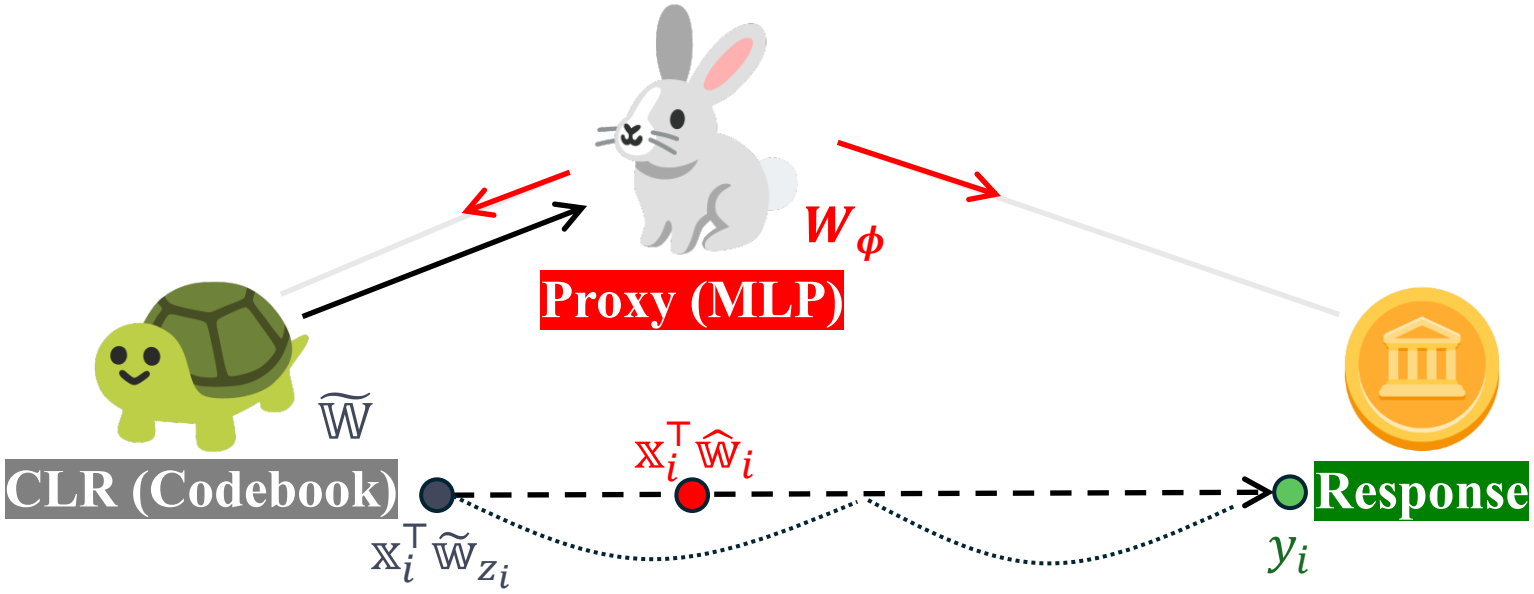


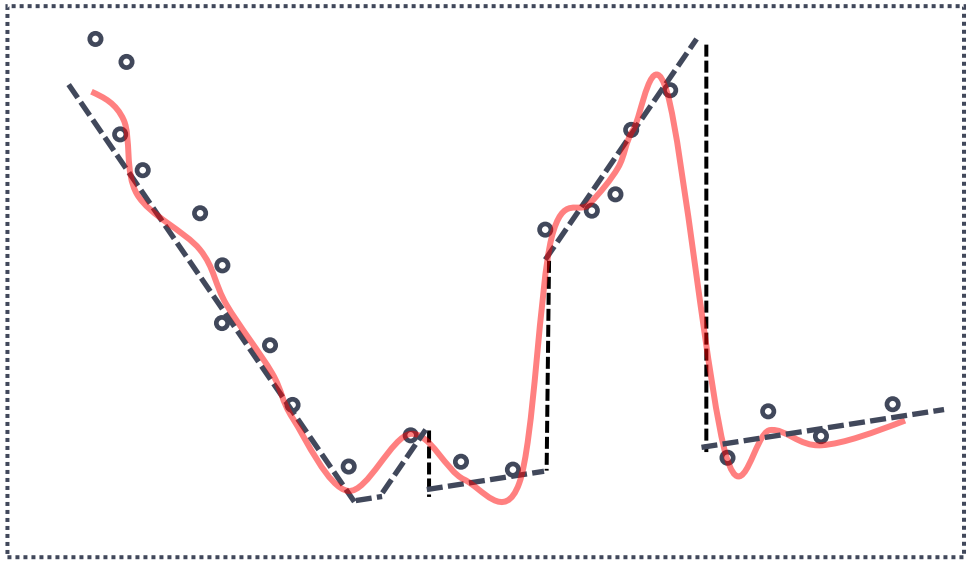


$$\phi^{(t+1)} \leftarrow \phi^{(t)} + \frac{2\eta(1+\lambda)}{N} \sum_{i=1}^{N} \left(\frac{y_i + \lambda \mathbb{x}_i^\top \tilde{\mathbb{w}}_{z_i}^{(t)}}{1+\lambda} - \mathbb{x}_i^\top \hat{\mathbb{w}}_i^{(t)} \right) (\nabla_\phi \hat{\mathbb{w}}_i^{(t)})^\top \mathbb{x}_i,$$

$$\tilde{\mathbb{w}}_j^{(t+1)} \leftarrow \tilde{\mathbb{w}}_j^{(t)} + \frac{2\eta(1+\lambda)}{N} \sum_{i \in S_j^{(t)}} \mathbb{x}_i \mathbb{x}_i^\top \left(\hat{\mathbb{w}}_i^{(t)} - \tilde{\mathbb{w}}_j^{(t)} \right), \quad \forall j \in [K].$$

Why does the Optimization stay Stable?

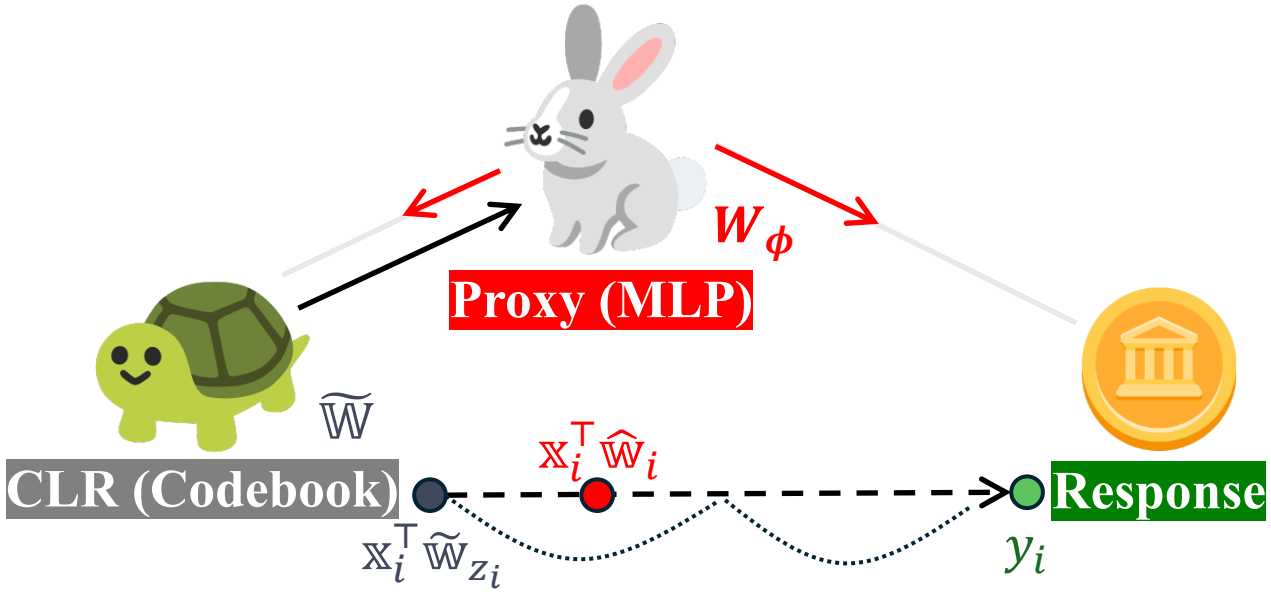


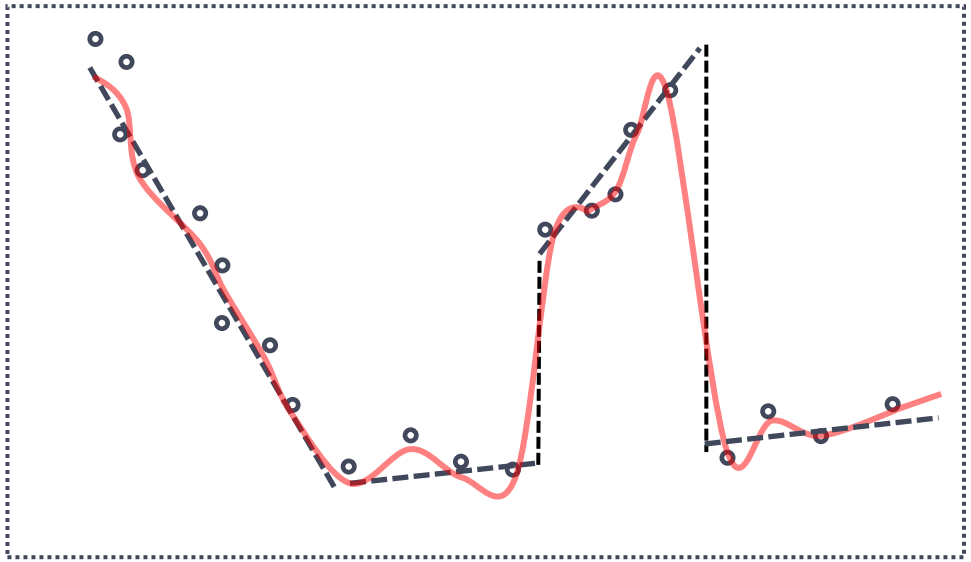


$$\phi^{(t+1)} \leftarrow \phi^{(t)} + \frac{2\eta(1+\lambda)}{N} \sum_{i=1}^{N} \left(\frac{y_i + \lambda \mathbb{x}_i^\top \tilde{\mathbb{w}}_{z_i}^{(t)}}{1+\lambda} - \mathbb{x}_i^\top \hat{\mathbb{w}}_i^{(t)} \right) (\nabla_\phi \hat{\mathbb{w}}_i^{(t)})^\top \mathbb{x}_i,$$

$$\tilde{\mathbb{w}}_j^{(t+1)} \leftarrow \tilde{\mathbb{w}}_j^{(t)} + \frac{2\eta(1+\lambda)}{N} \sum_{i \in S_j^{(t)}} \mathbb{x}_i \mathbb{x}_i^\top \left(\hat{\mathbb{w}}_i^{(t)} - \tilde{\mathbb{w}}_j^{(t)} \right), \quad \forall j \in [K].$$

Why does the Optimization stay Stable?

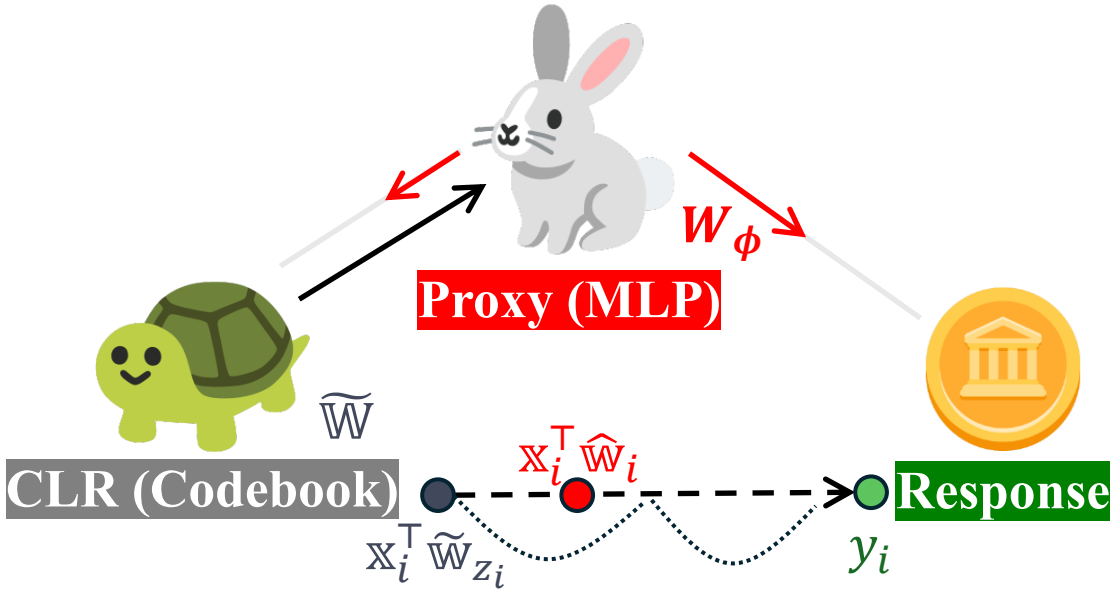


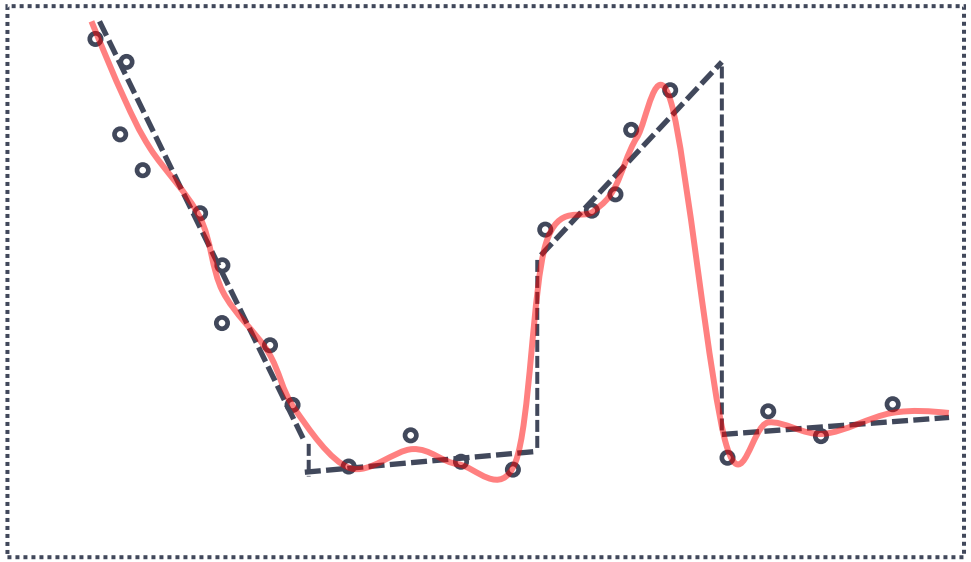


$$\phi^{(t+1)} \leftarrow \phi^{(t)} + \frac{2\eta(1+\lambda)}{N} \sum_{i=1}^{N} \left(\frac{y_i + \lambda \mathbb{x}_i^\top \tilde{\mathbb{w}}_{z_i}^{(t)}}{1+\lambda} - \mathbb{x}_i^\top \hat{\mathbb{w}}_i^{(t)} \right) (\nabla_\phi \hat{\mathbb{w}}_i^{(t)})^\top \mathbb{x}_i,$$

$$\tilde{\mathbb{w}}_j^{(t+1)} \leftarrow \tilde{\mathbb{w}}_j^{(t)} + \frac{2\eta(1+\lambda)}{N} \sum_{i \in S_j^{(t)}} \mathbb{x}_i \mathbb{x}_i^\top \left(\hat{\mathbb{w}}_i^{(t)} - \tilde{\mathbb{w}}_j^{(t)} \right), \quad \forall j \in [K].$$

Why does the Optimization stay Stable?





$$\phi^{(t+1)} \leftarrow \phi^{(t)} + \frac{2\eta(1+\lambda)}{N} \sum_{i=1}^{N} \left(\frac{y_i + \lambda \mathbb{x}_i^\top \tilde{\mathbb{w}}_{z_i}^{(t)}}{1+\lambda} - \mathbb{x}_i^\top \hat{\mathbb{w}}_i^{(t)} \right) (\nabla_\phi \hat{\mathbb{w}}_i^{(t)})^\top \mathbb{x}_i,$$

$$\tilde{\mathbb{w}}_j^{(t+1)} \leftarrow \tilde{\mathbb{w}}_j^{(t)} + \frac{2\eta(1+\lambda)}{N} \sum_{i \in S_j^{(t)}} \mathbb{x}_i \mathbb{x}_i^\top \left(\hat{\mathbb{w}}_i^{(t)} - \tilde{\mathbb{w}}_j^{(t)} \right), \quad \forall j \in [K].$$

Why does the Optimization stay Stable?

Proposition 3.1 (One-epoch descent). *Under Assumptions 1 and 2, suppose the cluster assignment $z^{(t)}$ remains fixed during epoch t. Define*

$$L_V := \max\Big\{(1+\lambda)L,\ (4+2\lambda)L_x^2\mathbb{X}_{\max}^2\Big\}.$$

For any step-size $0 < \eta \le 1/L_V$, the simultaneous gradient updates guarantee a strict decrease of the Lyapunov function:

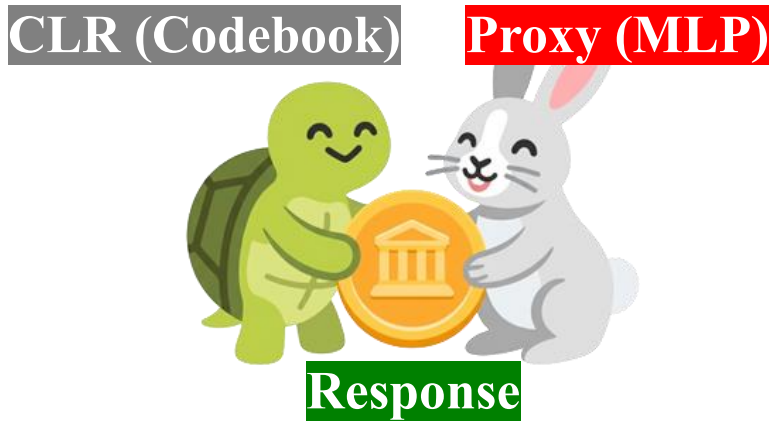
$$V_\lambda(\phi^{(t+1)}, \tilde{\mathbb{W}}^{(t+1)}) \le V_\lambda(\phi^{(t)}, \tilde{\mathbb{W}}^{(t)}) - \frac{\eta}{2}\Big(\|\phi^{(t+1)} - \phi^{(t)}\|^2 + \|\tilde{\mathbb{W}}^{(t+1)} - \tilde{\mathbb{W}}^{(t)}\|^2\Big).$$

This explicitly verifies that V_λ serves as a valid Lyapunov function, rigorously ensuring the convergence of the CG-CLR algorithm.

Theorem 3.2 (Linear convergence). *Let Assumptions 1–4 hold. Define*

$$\mu_V := \min\Big\{2m_x^2\mathbb{X}_{\min}^2,\ (1+\lambda)\mu\Big\}, \quad L_V := \max\Big\{(1+\lambda)L,\ (4+2\lambda)L_x^2\mathbb{X}_{\max}^2\Big\}, \quad q := \frac{L_V - \mu_V}{L_V + \mu_V}.$$

Gradient descent with the optimal step-size $\eta_\star = 2/(\mu_V + L_V)$ ensures that the stacked parameter vector $\theta = (\phi, \tilde{\mathbb{W}})$ converges linearly at the rate q, while the Lyapunov gap contracts at the rate q^2.



"The Proxy follows the interior point of Response and CLR, and CLR follows the Proxy"

Check out our paper & Join our poster!

😎 **Scan Me !**