
Supplementary Material - CCL: Causal-aware In-context Learning for Out-of-Distribution Generalization

Hoyoon Byun, Gyeongdeok Seo, Joonseong Kang, Taero Kim, Jihee Kim, Kyungwoo Song*
Department of Statistics and Data Science, Yonsei University
{hoyun.byun, gd.seo, doongsae, taero.kim, jihee_sta, kyungwoo.song}@yonsei.ac.kr

Contents

A	ELBO objective function	1
B	Proof of Theorem 3.3	4
C	Proof of Theorem 3.4	5
C.1	Preliminaries: population minimizer and gradient concentration	5
C.2	Supporting proposition: Condition of empirical Hessians	7
C.3	Main result: strictly better convergence for \mathcal{D}_c	9
D	Detailed experiment settings	12
D.1	Synthetic dataset	12
D.2	MGSM	12
D.2.1	ID / OOD dataset split	12
D.2.2	Task specific prompt	13
D.2.3	Fixed in-context samples	14
D.3	Advanced experiments	14

A ELBO objective function

To fit the model p_θ with the true source-domain distribution $p_{\theta^*}(x, y, t, e)$, we maximize the likelihood:

$$\max_{\theta} \mathbb{E}_{p_{\theta^*}(x, y, t, e)} [\log p_\theta(x, y, t, e)] \quad (1)$$

θ^* and θ denote the unknown true source-domain parameters and model parameters, respectively. Maximizing the likelihood is proportional to minimizing the KL divergence because the $\int p_{\theta^*}(x, y, t, e) \log p_\theta(x, y, t, e) dx dy dt de$ corresponds to the negative entropy, which is constant

*Corresponding author

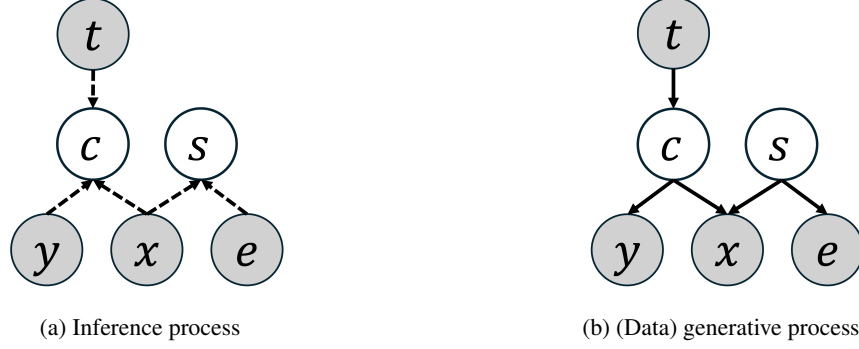


Figure 1: (a) Inference process in CCL. The variables x , y , t , and e are observed random variables represented as gray nodes, whereas c and s are latent random variables represented as white nodes. The dashed edges indicate the computation paths for inferring c and s . (b) Data generative process in CCL. We generate y , x , and e by reflecting their causal relationships with c and s along the causal paths. Since t serves as the root node of c , we define the conditional prior of c , $p_\theta(c | t)$. The edges represent paths where the computation path aligns with the generation path.

with respect to p_θ .

$$\begin{aligned}
 \text{KL}(p_{\theta^*}(x, y, t, e) || p_\theta(x, y, t, e)) &= \int p_{\theta^*}(x, y, t, e) \log \frac{p_{\theta^*}(x, y, t, e)}{p_\theta(x, y, t, e)} dx dy dt de \\
 &= \int p_{\theta^*}(x, y, t, e) \log p_{\theta^*}(x, y, t, e) dx dy dt de \\
 &\quad - \int p_{\theta^*}(x, y, t, e) \log p_\theta(x, y, t, e) dx dy dt de
 \end{aligned} \tag{2}$$

However, the marginal probability $p_\theta(x, y, t, e) = \int p_\theta(x, y, t, e, c, s) dc ds$ is intractable. To model a tractable distribution, we introduce the inference model $q_\phi(c, s | x, y, t, e)$, where ϕ represents the parameters of the tractable distribution. The lower bound of the log-likelihood function is derived as follows:

$$\begin{aligned}
 \log p_\theta(x, y, t, e) &= \log \int p_\theta(x, y, t, e, c, s) dc ds = \log \int q_\phi(c, s | x, y, t, e) \frac{p_\theta(x, y, t, e, c, s)}{q_\phi(c, s | x, y, t, e)} dc ds \\
 &= \log \mathbb{E}_{q_\phi(c, s | x, y, t, e)} \left[\frac{p_\theta(x, y, t, e, c, s)}{q_\phi(c, s | x, y, t, e)} \right] \\
 &\geq \mathbb{E}_{q_\phi(c, s | x, y, t, e)} \left[\log \frac{p_\theta(x, y, t, e, c, s)}{q_\phi(c, s | x, y, t, e)} \right]
 \end{aligned} \tag{3}$$

This lower bound of the log-likelihood function is termed the Evidence Lower Bound (ELBO). For simplicity, we denote this ELBO as L_{ELBO} . Maximizing the ELBO with respect to the source-domain distribution is equivalent to maximizing $\log p_\theta$. Furthermore, the ELBO and $\log p_\theta(x, y, t, e)$ are related as follows:

$$\begin{aligned}
 \text{ELBO} + \text{KL}(q_\phi(c, s | x, y, t, e) || p_\theta(c, s | x, y, t, e)) &= \mathbb{E}_{q_\phi(c, s | x, y, t, e)} \log p_\theta(x, y, t, e) + \mathbb{E}_{q_\phi(c, s | x, y, t, e)} \log \frac{p_\theta(c, s | x, y, t, e)}{q_\phi(c, s | x, y, t, e)} \\
 &\quad + \text{KL}(q_\phi(c, s | x, y, t, e) || p_\theta(c, s | x, y, t, e)) \\
 &= \log p_\theta(x, y, t, e)
 \end{aligned} \tag{4}$$

Since $\log p_\theta(x, y, t, e)$ is constant with respect to $q_\phi(c, s | x, y, t, e)$, maximizing the ELBO is equivalent to minimizing $\text{KL}(q_\phi(c, s | x, y, t, e) || p_\theta(c, s | x, y, t, e))$. Thus, maximizing the ELBO, i.e., optimizing θ and ϕ to obtain a tight lower bound of $\log p_\theta(x, y, t, e)$, implies that $q_\phi(c, s |$

x, y, t, e) approximates $p_\theta(c, s \mid x, y, t, e)$. Furthermore, it ensures that $p_\theta(x, y, t, e)$ aligns with the true source-domain distribution $p_{\theta^*}(x, y, t, e)$.

As θ^* is unknown, we instead use the observed data distribution in the source domain, $p_D(x, y, t, e)$, as the target for fitting:

$$\max_{\theta, \phi} \mathbb{E}_{p_D(x, y, t, e)} [L_{ELBO}] \quad (5)$$

The inference model $q_\phi(c, s \mid x, y, t, e)$ requires y as a condition. However, y is unavailable during test time. We reformulate the inference model by marginalizing y using the factorization, which relies on the conditional independence $y \perp (x, t, e, s) \mid c$ and the invariant causal mechanism $p_\theta(y \mid c)$:

$$q_\phi(c, s, y \mid x, t, e) = q_\phi(c, s \mid x, t, e) p_\theta(y \mid c)$$

Using Bayes' rule, the original inference model $q_\phi(c, s \mid x, y, t, e)$ (which conditions on y) can be expressed as:

$$q_\phi(c, s \mid x, y, t, e) = \frac{q_\phi(c, s, y \mid x, t, e)}{q_\phi(y \mid x, t, e)} = \frac{q_\phi(c, s \mid x, t, e) p_\theta(y \mid c)}{q_\phi(y \mid x, t, e)}$$

where we define $\Phi_{y|x,t,e} := q_\phi(y \mid x, t, e) = \int q_\phi(c, s \mid x, t, e) p_\theta(y \mid c) dc ds = \mathbb{E}_{q_\phi(c, s|x,t,e)} [p_\theta(y \mid c)]$.

Now, we substitute this back into the original L_{ELBO} objective. We also use the generative model factorization: $p_\theta(x, y, t, e, c, s) = p_\theta(x, t, e, c, s) p_\theta(y \mid c)$.

$$\begin{aligned} L_{ELBO} &= \mathbb{E}_{q_\phi(c, s|x,y,t,e)} \left[\log \frac{p_\theta(x, y, t, e, c, s)}{q_\phi(c, s \mid x, y, t, e)} \right] \\ &= \int q_\phi(c, s \mid x, y, t, e) \left[\log \frac{p_\theta(x, t, e, c, s) p_\theta(y \mid c)}{\frac{q_\phi(c, s|x,t,e) p_\theta(y|c)}{q_\phi(y|x,t,e)}} \right] dc ds \\ &= \int \frac{q_\phi(c, s \mid x, t, e) p_\theta(y \mid c)}{q_\phi(y \mid x, t, e)} \left[\log \frac{p_\theta(x, t, e, c, s)}{q_\phi(c, s \mid x, t, e)} + \log q_\phi(y \mid x, t, e) \right] dc ds \end{aligned}$$

We split this into two terms:

$$\begin{aligned} &\int \frac{q_\phi(c, s \mid x, t, e) p_\theta(y \mid c)}{q_\phi(y \mid x, t, e)} \log \frac{p_\theta(x, t, e, c, s)}{q_\phi(c, s \mid x, t, e)} dc ds \\ &+ \int \frac{q_\phi(c, s \mid x, t, e) p_\theta(y \mid c)}{q_\phi(y \mid x, t, e)} \log q_\phi(y \mid x, t, e) dc ds \end{aligned}$$

The second term simplifies, as $\log q_\phi(y \mid x, t, e)$ is constant w.r.t c, s :

$$\begin{aligned} &\int \frac{q_\phi(c, s \mid x, t, e) p_\theta(y \mid c)}{q_\phi(y \mid x, t, e)} \log q_\phi(y \mid x, t, e) dc ds \\ &= \frac{\log q_\phi(y \mid x, t, e)}{q_\phi(y \mid x, t, e)} \int q_\phi(c, s \mid x, t, e) p_\theta(y \mid c) dc ds \\ &= \frac{\log q_\phi(y \mid x, t, e)}{q_\phi(y \mid x, t, e)} \cdot q_\phi(y \mid x, t, e) \\ &= \log q_\phi(y \mid x, t, e) \end{aligned}$$

The first term is an expectation w.r.t $q_\phi(c, s \mid x, t, e)$:

$$\begin{aligned} &\int \frac{q_\phi(c, s \mid x, t, e) p_\theta(y \mid c)}{q_\phi(y \mid x, t, e)} \log \frac{p_\theta(x, t, e, c, s)}{q_\phi(c, s \mid x, t, e)} dc ds \\ &= \frac{1}{q_\phi(y \mid x, t, e)} \mathbb{E}_{q_\phi(c, s|x,t,e)} \left[p_\theta(y \mid c) \log \frac{p_\theta(x, t, e, c, s)}{q_\phi(c, s \mid x, t, e)} \right] \end{aligned}$$

Combining these terms, the L_{ELBO} for a single data point (x, y, t, e) is:

$$L_{ELBO} = \log q_\phi(y | x, t, e) + \frac{1}{q_\phi(y | x, t, e)} \mathbb{E}_{q_\phi(c, s | x, t, e)} \left[p_\theta(y | c) \log \frac{p_\theta(x, t, e, c, s)}{q_\phi(c, s | x, t, e)} \right]$$

Finally, taking the expectation over the data distribution $p_D(x, y, t, e)$ and substituting $\Phi_{y|x, t, e} = q_\phi(y | x, t, e)$, we arrive at the final objective function:

$$\begin{aligned} \max_{\theta, \phi} \mathbb{E}_{p_D(x, y, t, e)} [L_{ELBO}] &= \mathbb{E}_{p_D} \left[\log \Phi_{y|x, t, e} \right. \\ &\quad \left. + \frac{1}{\Phi_{y|x, t, e}} \mathbb{E}_{q_\phi(c, s | x, t, e)} [p_\theta(y | c) \log \frac{p_\theta(x, t, e, c, s)}{q_\phi(c, s | x, t, e)}] \right] \end{aligned}$$

B Proof of Theorem 3.3

Assumption B.1 (Sufficient column-space rank). We assume \mathcal{B}_2 has sufficient column-space rank (e.g., rank and column-space properties as discussed in the Remark on rank and offsetting) and the x noise difference $\Delta \varepsilon_x = \varepsilon_x^* - \varepsilon_x$ is small.

Assumption B.2 (Small noise difference). We assume y noise difference $\Delta \varepsilon_y = \varepsilon_y^* - \varepsilon_y$ is small, so that $|y^* - y| \approx |(w^*)^\top \Delta c|$. Also, we consider Δc such that it is not orthogonal to w^* .

Step 1: Showing that small $\|x^* - x\|$ does not imply small $\|c^* - c\|$. We have the following equations:

$$x^* = \mathcal{B}_1 c^* + \mathcal{B}_2 s^* + \varepsilon_x^*, \quad x = \mathcal{B}_1 c + \mathcal{B}_2 s + \varepsilon_x.$$

Hence, $x^* - x = \mathcal{B}_1(c^* - c) + \mathcal{B}_2(s^* - s) + (\varepsilon_x^* - \varepsilon_x)$. Suppose one wants $\|x^* - x\|$ to be *arbitrarily small*. Observe that even if $\|c^* - c\|$ is large, one can often choose s^* and s so that $\mathcal{B}_2(s^* - s)$ *offsets* $\mathcal{B}_1(c^* - c)$, thus making $x^* - x$ remain close to $\mathbf{0}$. Concretely, if $\Delta c := c^* - c$ is large, we can select

$$\Delta s := (s^* - s) \quad \text{such that} \quad \mathcal{B}_2 \Delta s \approx -\mathcal{B}_1 \Delta c.$$

Under Assumption B.1, $x^* - x$ can be made small. Thus, $\|x^* - x\| < \epsilon$ *does not* force $\|c^* - c\| \leq$ (something small); we can keep $\|c^* - c\|$ large by adjusting s^* and s .

Step 2: A large $\|c^* - c\|$ forces a large difference in y^* and y . Similarly,

$$y^* = (w^*)^\top c^* + \varepsilon_y^*, \quad y = (w^*)^\top c + \varepsilon_y.$$

Hence, $y^* - y = (w^*)^\top (c^* - c) + (\varepsilon_y^* - \varepsilon_y)$. If $\|c^* - c\|$ is large, then $(w^*)^\top \Delta c$ will also be large in magnitude, provided that $w^* \neq 0$ and Δc is not orthogonal to w^* . Formally, the magnitude is given by $|(w^*)^\top \Delta c| = \|w^*\| \|\Delta c\| \cos \theta$, where θ is the angle between w^* and Δc . If we consider Δc such that it is not orthogonal to w^* , then $|(w^*)^\top \Delta c|$ is proportional to $\|\Delta c\|$.

Step 3: Concluding there is no universal upper bound. From Step 1, even when $\|x^* - x\| < \epsilon$ is made arbitrarily small (by choosing s^* and s appropriately), one can choose c^* and c such that $\|\Delta c\| = \|c^* - c\|$ is large. Specifically, let $\kappa > 0$, and we choose Δc such that

$$\|\Delta c\| > \frac{\kappa}{\|\mathcal{B}_1\|_{\text{op}}}.$$

This choice for Δc is possible while maintaining $\|x^* - x\| < \epsilon$, due to the offsetting capability of $\mathcal{B}_2 \Delta s$. Now, we examine the consequence for $|y^* - y|$. Under Assumption B.2, we have the following inequality:

$$|(w^*)^\top \Delta c| = \|w^*\| \|\Delta c\| \cos \theta \geq \|w^*\| \cos \theta \frac{\kappa}{\|\mathcal{B}_1\|_{\text{op}}} = \kappa \frac{\|w^*\| \cos \theta}{\|\mathcal{B}_1\|_{\text{op}}}.$$

Therefore, there is no universal upper bound on $\|y^* - y\|$ solely from the condition $\|x^* - x\| \leq \epsilon$. As $\epsilon \rightarrow 0$, the difference $\|x^* - x\|$ in input space can vanish, yet $\|y^* - y\|$ may remain *unbounded* by selecting c^* sufficiently different from c and offsetting it with $s^* - s$. This completes the proof that a small x -distance does not prevent a large y -distance.

Remark on rank and offsetting. A key factor is that \mathcal{B}_2 must have enough degrees of freedom (rank) to offset $\mathcal{B}_1(c^* - c)$. The approximation, $\mathcal{B}_2\Delta s \approx -\mathcal{B}_1\Delta c$, can be achieved if the vector $-\mathcal{B}_1\Delta c$ lies within the column space of \mathcal{B}_2 , denoted as $\mathcal{C}(\mathcal{B}_2)$. A strong condition that guarantees exact offsetting is when $\mathcal{C}(\mathcal{B}_1) \subseteq \mathcal{C}(\mathcal{B}_2)$. When such offsetting can be achieved, $\|x^* - x\|$ can be made arbitrarily small despite a large $\|c^* - c\|$. Consequently, $\|y^* - y\| \approx \|(w^*)^\top \Delta c\|$ can become large.

C Proof of Theorem 3.4

Assumption C.1 (Zero-mean condition).

$$\begin{aligned} \mathbb{E}[c] &= \mathbb{E}[s] = \mathbb{E}[\varepsilon_x] = \mathbb{E}[\varepsilon_y] = 0, \\ \mathbb{E}[cs^\top] &= \mathbb{E}[c\varepsilon_x^\top] = \mathbb{E}[s\varepsilon_x^\top] = 0, \quad \mathbb{E}[c\varepsilon_y] = \mathbb{E}[s\varepsilon_y] = \mathbb{E}[\varepsilon_x\varepsilon_y] = 0. \end{aligned}$$

Our structural causal model assumes joint independence, which implies zero-mean cross terms.

Assumption C.2 (Subspace orthogonality).

$$\mathcal{B}_1^\top w^* = w^*, \quad \mathcal{B}_2^\top w^* = 0, \quad \Sigma_\varepsilon^{1/2} w^* = 0, \quad \text{where } \Sigma_\varepsilon = \mathbb{E}[\varepsilon_x \varepsilon_x^\top].$$

This implies that w^* purely reflects causal effects and is decoupled from the spurious parts or noise structure within x . Also, we assume $\mathbb{E}[xx^\top]$ is positive definite.

Assumption C.3 (No common null spaces). We denote \mathcal{N} for null space.

$$\mathcal{N}(\mathcal{B}_1 \mathbb{E}[\Delta c \Delta c^\top] \|\Delta c\| \leq \delta) \cap \mathcal{N}(\mathcal{B}_2 \Sigma_s \mathcal{B}_2^\top) \cap \mathcal{N}(\Sigma_\varepsilon) \cap \mathbf{v}^\perp = \{\mathbf{0}\}.$$

Here, $\Sigma_s = \mathbb{E}[ss^\top]$, $\Sigma_\varepsilon = \mathbb{E}[\varepsilon_x \varepsilon_x^\top]$, $\mathbf{v} = \mathcal{B}_1 c^*$.

Assumption C.4 (Bounded norm and neighborhood radius). We denote $r_{(N)} := \max_{i \in \mathcal{D}_x} \|x_i - x^*\|$. Then, we assume for sufficiently large N and arbitrary small δ ,

$$(\|x^*\| + r_{(N)})^2 < C_{l,u}(\Lambda_{c,\min}^* - \mathcal{K}\delta),$$

where $\Lambda_{c,\min}^*$ and \mathcal{K} are defined in Appendix C.2. $C_{l,u}$ is a positive constant less than or equal to 1 defined in Proof of Claim 1. The norm will be small if $\mathcal{B}_1 c^* \approx -\mathcal{B}_2 s^*$ which means x^* has lack of information on c^* . Also, we assume $\text{tr}(\mathbb{E}[xx^\top(y - x^\top w^*)^2]) > 0$, where x and y are random variables from \mathcal{D}_x .

Assumption C.5 (Cosine alignment). We assume the cosines of the angles between x^* and $\Delta w_c, \Delta w_x$ are bounded away from zero and approximately equal, where $\Delta w_c = w_c^{(M)} - w^*$ and $\Delta w_x = w_x^{(M)} - w^*$. In other words, there exists $\rho \in (0, 1]$ such that

$$|(\Delta w_x)^\top x^*| - |(\Delta w_c)^\top x^*| \geq \rho \|x^*\| (\|\Delta w_x\| - \|\Delta w_c\|).$$

Assumption C.6 (Test offset margin). There exists $\mathcal{J} > 0$ such that

$$2|\mathcal{R}| \leq \rho \|x^*\| \alpha - \mathcal{J}.$$

Here, $\mathcal{R} = (w^*)^\top x^* - (w^*)^\top c^* - \varepsilon_y^*$, ρ , and α are defined in Appendix C.3.

C.1 Preliminaries: population minimizer and gradient concentration

Lemma C.7 (Population minimizer is w^*). *The unique minimizer of the population MSE loss*

$$\mathcal{L}(w) := \mathbb{E}[(w^\top x - y)^2]$$

is precisely w^* . Concretely,

$$w^* = \arg \min_{w \in \mathbb{R}^d} \mathcal{L}(w).$$

Proof of Lemma C.7

Expanding the population risk, $\mathcal{L}(w) = \mathbb{E}[(w^\top x - y)^2] = \mathbb{E}[w^\top x - (w^*)^\top c - \varepsilon_y]^2$, together with $x = \mathcal{B}_1 c + \mathcal{B}_2 s + \varepsilon_x$. The derivative $\nabla_w \mathcal{L}(w)$ vanishes exactly if w matches w^* in the subspace spanned by \mathcal{B}_1 . We show that w^* is the unique solution. A standard derivative calculation for MSE shows

$$\nabla_w \mathcal{L}(w) = 2 \left(\mathbb{E}[xx^\top]w - \mathbb{E}[xy] \right).$$

Hence, the minimizer must satisfy

$$\nabla_w \mathcal{L}(w) = 0 \iff \mathbb{E}[xx^\top]w = \mathbb{E}[xy].$$

This is the normal equation for linear regression. Given $y = (w^*)^\top c + \varepsilon_y$ and $x = \mathcal{B}_1 c + \mathcal{B}_2 s + \varepsilon_x$, by Assumption C.1

$$\mathbb{E}[xy] = \mathcal{B}_1 \mathbb{E}[cc^\top](w^*).$$

All other terms vanish in expectation. Similarly,

$$xx^\top = (\mathcal{B}_1 c + \mathcal{B}_2 s + \varepsilon_x)(\mathcal{B}_1 c + \mathcal{B}_2 s + \varepsilon_x)^\top.$$

Taking expectation (and discarding cross terms) yields

$$\mathbb{E}[xx^\top] = \mathcal{B}_1 \mathbb{E}[cc^\top] \mathcal{B}_1^\top + \mathcal{B}_2 \mathbb{E}[ss^\top] \mathcal{B}_2^\top + \mathbb{E}[\varepsilon_x \varepsilon_x^\top].$$

Here, the first block is the causal contribution, the second block is spurious, and the last is noise. We define $\Sigma_x = \mathbb{E}[xx^\top]$, $\Sigma_c = \mathbb{E}[cc^\top]$, $\Sigma_s = \mathbb{E}[ss^\top]$, $\Sigma_\varepsilon = \mathbb{E}[\varepsilon_x \varepsilon_x^\top]$. We get

$$\left(\mathcal{B}_1 \Sigma_c \mathcal{B}_1^\top + \mathcal{B}_2 \Sigma_s \mathcal{B}_2^\top + \Sigma_\varepsilon \right) w = \mathcal{B}_1 \Sigma_c w^*. \quad (6)$$

By Assumption C.2, $(\mathcal{B}_2 \Sigma_s \mathcal{B}_2^\top + \Sigma_\varepsilon)w^* = 0$. The solution (6) is $w = w^*$. A linear MSE objective is strictly convex if Σ_x is positive-definite in the relevant directions. Hence, it has exactly one minimizer. We conclude w^* is the unique minimizer.

Lemma C.8 (Uniform gradient concentration). *Let $\hat{L}_N(\cdot; \mathcal{D})$ be the empirical MSE for samples (x_i, y_i) drawn i.i.d. from a sub-Gaussian linear model (Assumption 3.1). Then for any compact set $\mathcal{W} \subset \mathbb{R}^d$, with probability at least $1 - \delta_{\text{tail}}$,*

$$\sup_{w \in \mathcal{W}} \left\| \nabla_w \hat{L}_N(w; \mathcal{D}) - \nabla_w \mathcal{L}(w) \right\| \leq \varepsilon_{\text{grad}}.$$

Here $\varepsilon_{\text{grad}} = O\left(\sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}}\right)$. By applying standard sub-Gaussian matrix Bernstein or Hanson–Wright inequalities to handle the $\{x_i x_i^\top\}$ terms and linear forms, this guarantees uniform closeness of empirical to population gradients in a finite ball. See, e.g., [1] for details on sub-Gaussian concentration.

Proof of Lemma C.8

We consider an MSE objective of the form

$$\hat{L}_N(w; \mathcal{D}) = \frac{1}{N} \sum_{i \in \mathcal{D}} (w^\top x_i - y_i)^2.$$

For a demonstration set \mathcal{D} of size N , we run M steps of gradient descent with step size η : $m = 0, \dots, M-1$,

$$w^{(m+1)} = w^{(m)} - \eta \nabla_w \hat{L}_N(w^{(m)}; \mathcal{D}) \quad (7)$$

We denote the final iteration by $w^{(M)}$ and $\nabla := \nabla_w$. In D_c , the final iterate returns $w_c^{(M)}$; for D_x , it returns $w_x^{(M)}$.

Step 1: Rewrite the gradient difference. For each $w \in \mathcal{W}$,

$$(\nabla \hat{L}_N(w)) - (\nabla \mathcal{L}(w)) = 2 \left[\underbrace{\left(\frac{1}{N} \sum_{i \in \mathcal{D}} x_i x_i^\top \right) - \mathbb{E}[x x^\top]}_{=: A_N} \right] w - 2 \left[\underbrace{\left(\frac{1}{N} \sum_{i \in \mathcal{D}} x_i y_i \right) - \mathbb{E}[x y]}_{=: b_N} \right].$$

Hence

$$\|\nabla \hat{L}_N(w) - \nabla \mathcal{L}(w)\| = 2\|A_N w - b_N\|.$$

Thus we require

$$\sup_{w \in \mathcal{W}} \|A_N w - b_N\| \text{ to be small, with high probability.}$$

Step 2: Matrix concentration for A_N . Define $A_N = \frac{1}{N} \sum_{i=1}^N x_i x_i^\top - \mathbb{E}[x x^\top]$. Under sub-Gaussian assumptions on x_i , the operator norm $\|A_N\|_{\text{op}}$ typically satisfies

$$\|A_N\|_{\text{op}} \leq C_1 \sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}}$$

with probability $\geq 1 - \delta_{\text{tail}}/2$ for some constant C_1 . This is a standard result from matrix Bernstein or Hanson–Wright inequalities ([1]). Hence, for $w \in \mathcal{W}$ with $\bar{W} = \sup_{w \in \mathcal{W}} \|w\|$, we get

$$\sup_{\|w\| \leq \bar{W}} \|A_N w\| \leq \|A_N\|_{\text{op}} \bar{W} \leq C_1 \bar{W} \sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}} \quad (\text{with probability } \geq 1 - \delta_{\text{tail}}/2).$$

Step 3: Vector concentration for b_N . Define $b_N = \left(\frac{1}{N} \sum_{i=1}^N x_i y_i \right) - \mathbb{E}[x y]$. Since x_i and y_i remain sub-Gaussian (or sub-exponential) under Assumption 3.1 with independence assumptions, each coordinate of $x_i y_i$ has tail bounds. A vector Bernstein (or Hoeffding) bound says for some constant C_2

$$\|b_N\| \leq C_2 \sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}} \quad (\text{with probability } \geq 1 - \delta_{\text{tail}}/2).$$

Step 4: Combine via union bound. By union bound, with probability at least $1 - \delta_{\text{tail}}$, both bounds above hold. Then for all $w \in \mathcal{W}$,

$$\|A_N w - b_N\| \leq \|A_N w\| + \|b_N\| \leq O\left(\sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}}\right).$$

Finally,

$$\|\nabla \hat{L}_N(w) - \nabla \mathcal{L}(w)\| = 2\|A_N w - b_N\| \leq O\left(\sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}}\right).$$

Since $w \in \mathcal{W}$ was arbitrary, we obtain

$$\sup_{w \in \mathcal{W}} \|\nabla \hat{L}_N(w; \mathcal{D}) - \nabla \mathcal{L}(w)\| \leq O\left(\sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}}\right).$$

This completes the proof of Lemma C.8.

C.2 Supporting proposition: Condition of empirical Hessians

Proposition C.9 (Conditioning of empirical Hessians). *Let the empirical covariance be $\Gamma(\mathcal{D}) = \frac{1}{N} \sum_{i \in \mathcal{D}} x_i x_i^\top$. Then, the empirical Hessian is $\nabla^2 \hat{L}_N(w; \mathcal{D}) = 2\Gamma(\mathcal{D})$. For sufficiently large N , with high probability (at least $1 - \delta_{\text{tail}}$):*

- The minimum eigenvalue of the empirical Hessian for \mathcal{D}_c , denoted by $\hat{\lambda}_c = \lambda_{\min}(\nabla^2 \hat{L}_N(\cdot; \mathcal{D}_c))$, is bounded below by a positive value: $\hat{\lambda}_c \geq 2(\Lambda_{c,\min}^* - \mathcal{K}\delta)$, where $\Lambda_{c,\min}^* > 0$.

- The minimum eigenvalue of the empirical Hessian for \mathcal{D}_x , $\hat{\lambda}_x = \lambda_{\min}(\nabla^2 \hat{L}_N(\cdot; \mathcal{D}_x))$, can be smaller than $\hat{\lambda}_c$.
- The maximum eigenvalue of $\Gamma(\mathcal{D}_x)$ can be smaller than $\Lambda_{c,\min}^* - \mathcal{K}\delta$.

Proof of Proposition C.9

The proof follows three main steps: first, we analyze the population covariance for data sampled under c -similarity; second, we connect this to the empirical covariance for \mathcal{D}_c via concentration inequalities; third, we contrast this with the situation for \mathcal{D}_x .

Step 1: Population covariance conditioned on c -similarity. Let $\Sigma_\delta = \mathbb{E}[xx^\top | \|\Delta c\| \leq \delta]$. Since $x = \mathcal{B}_1 c + \mathcal{B}_2 s + \varepsilon_x$ and $\|\Delta c\| \leq \delta$, x consistently contains a component $\mathcal{B}_1 c^*$. Let $\mathbf{v} := \mathcal{B}_1 c^*$, then for all samples with $\|\Delta c\| \leq \delta$,

$$x = \mathbf{v} + \mathbf{e}, \text{ where } \mathbf{e} = \mathcal{B}_1 \Delta c + \mathcal{B}_2 s + \varepsilon_x, \\ xx^\top = (\mathbf{v} + \mathbf{e})(\mathbf{v} + \mathbf{e})^\top = \mathbf{v}\mathbf{v}^\top + \mathbf{v}\mathbf{e}^\top + \mathbf{e}\mathbf{v}^\top + \mathbf{e}\mathbf{e}^\top.$$

Taking the conditional expectation yields

$$\Sigma_\delta = \mathbb{E}[xx^\top | \|\Delta c\| \leq \delta] = \mathbf{v}\mathbf{v}^\top + \mathbb{E}[\mathbf{v}\mathbf{e}^\top + \mathbf{e}\mathbf{v}^\top + \mathbf{e}\mathbf{e}^\top | \|\Delta c\| \leq \delta].$$

The expectation of the remaining terms involves \mathbf{e} . The magnitude of $\mathcal{B}_1 \Delta c$ within \mathbf{e} is controlled by δ .

$$\mathbf{z}^\top \Sigma_\delta \mathbf{z} \geq \mathbf{z}^\top \mathbf{v}\mathbf{v}^\top \mathbf{z} - \mathbf{z}^\top \mathbb{E}[\mathbf{v}\mathbf{e}^\top + \mathbf{e}\mathbf{v}^\top + \mathbf{e}\mathbf{e}^\top | \|\Delta c\| \leq \delta] \mathbf{z}, \text{ for any unit vector } \mathbf{z}.$$

The minimum eigenvalue of Σ_δ is denoted by $\lambda_{\min}(\Sigma_\delta) = \min_{\|\mathbf{z}\|=1} \mathbf{z}^\top \Sigma_\delta \mathbf{z}$.

$$\begin{aligned} \mathbb{E}[\mathbf{e} | \|\Delta c\| \leq \delta] &= \mathbb{E}[\mathcal{B}_1 \Delta c | \|\Delta c\| \leq \delta] + \mathbb{E}[\mathcal{B}_2 s | \|\Delta c\| \leq \delta] + \mathbb{E}[\varepsilon_x | \|\Delta c\| \leq \delta] \\ &= \mathcal{B}_1 \mathbb{E}[\Delta c | \|\Delta c\| \leq \delta] + \mathcal{B}_2 \mathbb{E}[s | \|\Delta c\| \leq \delta] + \mathbb{E}[\varepsilon_x | \|\Delta c\| \leq \delta] \\ &= \mathcal{B}_1 \mathbb{E}[\Delta c | \|\Delta c\| \leq \delta] \quad \dots \mathbb{E}[s] = \mathbb{E}[\varepsilon_x] = 0. \end{aligned}$$

The norm is upper-bounded.

$$\begin{aligned} \|\mathbb{E}[\mathbf{e} | \|\Delta c\| \leq \delta]\| &= \|\mathcal{B}_1 \mathbb{E}[\Delta c | \|\Delta c\| \leq \delta]\| \leq \|\mathcal{B}_1\|_{\text{op}} \cdot \|\mathbb{E}[\Delta c | \|\Delta c\| \leq \delta]\| \\ &\leq \|\mathcal{B}_1\|_{\text{op}} \cdot \mathbb{E}[\|\Delta c\| | \|\Delta c\| \leq \delta] \quad \dots \text{Jensen's inequality} \\ &\leq \|\mathcal{B}_1\|_{\text{op}} \cdot \delta. \end{aligned}$$

We denote $\mu_{\mathbf{e}} := \mathbb{E}[\mathbf{e} | \|\Delta c\| \leq \delta]$, $\sigma_{\min}(\mathcal{B}_2)^2 := \lambda_{\min}(\mathcal{B}_2^\top \mathcal{B}_2)$, $\Sigma_{\mathbf{e}} := \mathbb{E}[\mathbf{e}\mathbf{e}^\top]$.

$$\begin{aligned} \mathbf{z}^\top \Sigma_\delta \mathbf{z} &= (\mathbf{z}^\top \mathbf{v})^2 + 2(\mathbf{z}^\top \mathbf{v})(\mathbf{z}^\top \mu_{\mathbf{e}}) + \mathbf{z}^\top \mathbb{E}[\mathbf{e}\mathbf{e}^\top | \|\Delta c\| \leq \delta] \mathbf{z} \\ &\geq (\mathbf{z}^\top \mathbf{v})^2 + \mathbf{z}^\top \mathbb{E}[\mathbf{e}\mathbf{e}^\top | \|\Delta c\| \leq \delta] \mathbf{z} - 2\|\mathbf{v}\|\|\mu_{\mathbf{e}}\| \\ &= \mathbf{z}^\top (\mathbf{v}\mathbf{v}^\top + \mathbb{E}[\mathbf{e}\mathbf{e}^\top | \|\Delta c\| \leq \delta]) \mathbf{z} - 2\|\mathbf{v}\|\|\mu_{\mathbf{e}}\| \\ &\geq \mathbf{z}^\top (\mathbf{v}\mathbf{v}^\top + \mathbb{E}[\mathbf{e}\mathbf{e}^\top | \|\Delta c\| \leq \delta]) \mathbf{z} - 2\|\mathbf{v}\|\|\mathcal{B}_1\|_{\text{op}}\delta \\ &\geq \underbrace{\lambda_{\min}(\mathbf{v}\mathbf{v}^\top + \mathbb{E}[\mathbf{e}\mathbf{e}^\top | \|\Delta c\| \leq \delta])}_{=:\Lambda_{c,\min}} - 2\|\mathbf{v}\|\|\mathcal{B}_1\|_{\text{op}}\delta. \end{aligned}$$

Under Assumption C.3, $\Lambda_{c,\min} > 0$. Therefore,

$$\lambda_{\min}(\Sigma_\delta) \geq \Lambda_{c,\min} - \mathcal{K}\delta, \quad \mathcal{K} := 2\|\mathbf{v}\|\|\mathcal{B}_1\|_{\text{op}}.$$

Step 2: Sample covariance concentration. The empirical covariance for \mathcal{D}_c is $\Gamma(\mathcal{D}_c) = \frac{1}{N} \sum_{i \in \mathcal{D}_c} x_i x_i^\top$. Under standard sub-Gaussian assumptions for ε_x , applying matrix Bernstein (or Hanson-Wright) inequalities concludes for some $C_\Gamma > 0$

$$\|\Gamma(\mathcal{D}_c) - \Sigma_\delta\|_{\text{op}} \leq C_\Gamma \left(\sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}} \right) \quad \text{with probability at least } 1 - \delta_{\text{tail}}.$$

The minimum eigenvalue of the empirical covariance for \mathcal{D}_c is bounded below as follows:

$$\begin{aligned}\lambda_{\min}(\Gamma(\mathcal{D}_c)) &\geq \lambda_{\min}(\Sigma_\delta) - \|\Gamma(\mathcal{D}_c) - \Sigma_\delta\|_{\text{op}} && \text{By Weyl's inequality} \\ &\geq \Lambda_{c,\min} - \mathcal{K}\delta - C_\Gamma \left(\sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}} \right) && \text{From Step 1.}\end{aligned}$$

Let $\Lambda_{c,\min}^* = \Lambda_{c,\min} - C_\Gamma \left(\sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}} \right)$. Then the minimum eigenvalue of the empirical Hessian satisfies the following inequality.

$$\lambda_{\min}(\nabla^2 \hat{L}_N(\cdot; \mathcal{D}_c)) = 2\lambda_{\min}(\Gamma(\mathcal{D}_c)) \geq 2(\Lambda_{c,\min}^* - \mathcal{K}\delta) \quad \text{with probability at least } 1 - \delta_{\text{tail}}.$$

Step 3: Ill-conditioning of $\Gamma(\mathcal{D}_x)$. The dataset \mathcal{D}_x consists of points that are nearest to x^* . Let $r_{(N)} := \max_{i \in \mathcal{D}_x} \|x_i - x^*\|$ and \bar{x} as the sample mean of data from \mathcal{D}_x . We denote $\Sigma_x := \frac{1}{N} \sum_{i \in \mathcal{D}_x} (x_i - \bar{x})(x_i - \bar{x})^\top$, so that $\Gamma(\mathcal{D}_x) = \Sigma_x + \bar{x}\bar{x}^\top$.

$$\lambda_{\min}(\Gamma(\mathcal{D}_x)) \leq \min_{\|\mathbf{z}\|=1} \mathbf{z}^\top \Gamma(\mathcal{D}_x) \mathbf{z}$$

Then, when we select \mathbf{z} such that $\mathbf{z} \perp \bar{x}$ (for $d \geq 2$ or $\bar{x} = 0$),

$$\begin{aligned}\lambda_{\min}(\Gamma(\mathcal{D}_x)) &\leq \mathbf{z}^\top \Gamma(\mathcal{D}_x) \mathbf{z} = \mathbf{z}^\top \Sigma_x \mathbf{z} \\ &\leq \text{tr}(\Sigma_x) = \frac{1}{N} \sum_{i \in \mathcal{D}_x} \|x_i - \bar{x}\|^2 \\ &\leq \frac{1}{N} \sum_{i \in \mathcal{D}_x} \|x_i - x^*\|^2 \leq r_{(N)}^2.\end{aligned}$$

Under Assumption C.4, with high probability for sufficiently large N and arbitrary small δ ,

$$\lambda_{\min}(\Gamma(\mathcal{D}_x)) < \Lambda_{c,\min}^* - \mathcal{K}\delta.$$

Thus, $\hat{\lambda}_x = 2\lambda_{\min}(\Gamma(\mathcal{D}_x))$ can be significantly smaller than $\hat{\lambda}_c$ and potentially approach zero. This implies that the empirical loss $\hat{L}_N(\cdot; \mathcal{D}_x)$ lacks robust, strong convexity in the directions essential for learning w^* . When it comes to the maximum eigenvalue, we denote $\lambda_{\max}(\Gamma(\mathcal{D}_x))$ as the maximum eigenvalue of $\Gamma(\mathcal{D}_x)$, which has the following upper bound.

$$\lambda_{\max}(\Gamma(\mathcal{D}_x)) \leq \frac{1}{N} \sum_{i \in \mathcal{D}_x} \|x_i\|^2 \leq (\|x^*\| + r_{(N)})^2.$$

Under Assumption C.4, with high probability for sufficiently large N and arbitrary small δ ,

$$\lambda_{\max}(\Gamma(\mathcal{D}_x)) \leq (\|x^*\| + r_{(N)})^2 < \Lambda_{c,\min}^* - \mathcal{K}\delta.$$

This inequality is a key to prove Claim 1.

C.3 Main result: strictly better convergence for \mathcal{D}_c

In this section, we prove the two key claims:

Claim 1. *With high probability, for sufficiently large M gradient-descent steps and $\alpha > 0$,*

$$\|w_c^{(M)} - w^*\| \leq \|w_x^{(M)} - w^*\| - \alpha.$$

Hence the parameter $w_c^{(M)}$ from the causal-similar set is strictly closer to w^ .*

Claim 2. *Consequently, the test error is also strictly smaller, for some $\alpha' > 0$.*

$$|(w_c^{(M)})^\top x^* - y^*| \leq |(w_x^{(M)})^\top x^* - y^*| - \alpha'.$$

Proof of Claim 1 (Parameter proximity)

$w_c^{(M)}$ and $w_x^{(M)}$ are obtained after M steps of gradient descent on the empirical loss functions $\widehat{L}_N(w; \mathcal{D}_c)$ and $\widehat{L}_N(w; \mathcal{D}_x)$, respectively. We denote an empirical minimizer $\hat{w}_{\mathcal{D}}$ as $\hat{w}_{\mathcal{D}} := \arg \min_w \widehat{L}_N(w; \mathcal{D})$. For a sufficiently large number of iterations M , $w_c^{(M)}$ will be close to the empirical minimizer \hat{w}_c of $\widehat{L}_N(w; \mathcal{D}_c)$, and $w_x^{(M)}$ will be close to \hat{w}_x , the empirical minimizer of $\widehat{L}_N(w; \mathcal{D}_x)$. Thus, the core of the argument lies in comparing the distances $\|\hat{w}_c - w^*\|$ and $\|\hat{w}_x - w^*\|$.

The empirical minimizer in linear model with the loss of mean squared error is given by $\hat{w}_{\mathcal{D}} = (\sum_{i \in \mathcal{D}} x_i x_i^\top)^{-1} (\sum_{i \in \mathcal{D}} x_i y_i) = (\Gamma(\mathcal{D}))^{-1} \widehat{\mathbb{E}}_{\mathcal{D}}[xy]$. The error $\hat{w}_{\mathcal{D}} - w^*$ can be expressed as:

$$\hat{w}_{\mathcal{D}} - w^* = (\Gamma(\mathcal{D}))^{-1} \left(\frac{1}{N} \sum_{i \in \mathcal{D}} x_i y_i - \Gamma(\mathcal{D}) w^* \right) = (\Gamma(\mathcal{D}))^{-1} \left(\frac{1}{N} \sum_{i \in \mathcal{D}} x_i (y_i - x_i^\top w^*) \right).$$

Note that $\mathbb{E}[x_i (y_i - x_i^\top w^*)] = \mathbb{E}[x_i y_i] - \mathbb{E}[x_i x_i^\top] w^* = 0$ by the definition of w^* from the normal equations (see proof of Lemma C.7). The term $Z_{\mathcal{D}} = \frac{1}{N} \sum_{i \in \mathcal{D}} x_i (y_i - x_i^\top w^*)$ is a sample average of zero-mean random vectors. Thus, $\|\hat{w}_{\mathcal{D}} - w^*\| \leq \|(\Gamma(\mathcal{D}))^{-1}\|_{\text{op}} \|Z_{\mathcal{D}}\|$. The operator norm $\|(\Gamma(\mathcal{D}))^{-1}\|_{\text{op}} = 1/\lambda_{\min}(\Gamma(\mathcal{D}))$.

From Proposition C.9:

- For \mathcal{D}_c : $\lambda_{\min}(\Gamma(\mathcal{D}_c)) \geq \Lambda_{c,\min}^* - \mathcal{K}\delta$. So, $\|(\Gamma(\mathcal{D}_c))^{-1}\|_{\text{op}} \leq \frac{1}{\Lambda_{c,\min}^* - \mathcal{K}\delta}$.
- For \mathcal{D}_x : $\lambda_{\min}(\Gamma(\mathcal{D}_x))$ can be significantly smaller than $\Lambda_{c,\min}^*$. Let $\lambda_{x,\text{eff}} = \lambda_{\min}(\Gamma(\mathcal{D}_x))$.

$$\begin{aligned} \nabla \widehat{L}_N(w; \mathcal{D}) &= 2 \left(\Gamma(\mathcal{D}) w - \frac{1}{N} \sum_{i \in \mathcal{D}} x_i y_i \right), \quad \Gamma(\mathcal{D}) = \frac{1}{N} \sum_{i \in \mathcal{D}} x_i x_i^\top. \\ \nabla \widehat{L}_N(w^*; \mathcal{D}) - \nabla L(w^*) &= 2 \left(\frac{1}{N} \sum_{i \in \mathcal{D}} x_i (x_i^\top w^* - y_i) \right) - 0 = -2Z_{\mathcal{D}}. \end{aligned}$$

By Lemma C.8,

$$\begin{aligned} \|2Z_{\mathcal{D}}\| &= \|\nabla \widehat{L}_N(w^*; \mathcal{D}) - \nabla L(w^*)\| \leq \varepsilon_{\text{grad}}, \\ \|Z_{\mathcal{D}}\| &\leq \frac{1}{2} \varepsilon_{\text{grad}} = O\left(\sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}}\right). \end{aligned}$$

Therefore, there exists a constant C_u for $\|Z_{\mathcal{D}_c}\|$ and $\|Z_{\mathcal{D}_x}\|$, with probability at least $1 - \delta_{\text{tail}}$,

$$\|Z_{\mathcal{D}_c}\| \leq C_u S_N \quad \text{and} \quad \|Z_{\mathcal{D}_x}\| \leq C_u S_N.$$

Here, $S_N = \sqrt{\frac{\log(1/\delta_{\text{tail}})}{N}}$. The distances have the following upper bounds.

$$\begin{aligned} \|\hat{w}_c - w^*\| &\leq \frac{1}{\Lambda_{c,\min}^* - \mathcal{K}\delta} \|Z_{\mathcal{D}_c}\|, \\ \|\hat{w}_x - w^*\| &\leq \frac{1}{\lambda_{x,\text{eff}}} \|Z_{\mathcal{D}_x}\|. \end{aligned}$$

We also utilize the following lower bound.

$$\frac{1}{\lambda_{\max}(\Gamma(\mathcal{D}_x))} \|Z_{\mathcal{D}_x}\| \leq \|\hat{w}_x - w^*\|.$$

Under the positive trace condition in Assumption C.4, we utilize Paley–Zygmund since $\|Z_{\mathcal{D}_x}\|$ is a positive random variable with finite 4th moment. There exists a constant c_l and $\delta_0 \in (0, 1)$ such that, for all sufficiently large N

$$P\left(\|Z_{\mathcal{D}_x}\| \geq \frac{c_l}{\sqrt{N}}\right) \geq 1 - \delta_0.$$

Fixing $\delta_{\text{tail}} = \delta_0$ yields $\|Z_{\mathcal{D}_x}\| \geq c_l S_N$.

$$\|\hat{w}_x - w^*\| - \|\hat{w}_c - w^*\| \geq \underbrace{\left(\frac{c_l}{\lambda_{\max}(\Gamma(\mathcal{D}_x))} - \frac{C_u}{\Lambda_{c,\min}^* - \mathcal{K}\delta} \right)}_{=: \alpha} S_N.$$

We have following inequality.

$$\frac{c_l}{\lambda_{\max}(\Gamma(\mathcal{D}_x))} - \frac{C_u}{\Lambda_{c,\min}^* - \mathcal{K}\delta} > \frac{c_l}{(\|x^*\| + r_{(N)})^2} - \frac{C_u}{\Lambda_{c,\min}^* - \mathcal{K}\delta}.$$

Under Assumption C.4, we denote $C_{l,u} = c_l/C_u$.

$$(\|x^*\| + r_{(N)})^2 < \frac{c_l}{C_u} (\Lambda_{c,\min}^* - \mathcal{K}\delta).$$

Then, there exists $\alpha > 0$. This leads to $\|\hat{w}_c - w^*\|$ being smaller than $\|\hat{w}_x - w^*\|$. The difference can be characterized by some $\alpha > 0$. The difference $\|\hat{w}_x - w^*\| - \|\hat{w}_c - w^*\|$ gives the margin α . For sufficiently large M , $w_c^{(M)} \approx \hat{w}_c$ and $w_x^{(M)} \approx \hat{w}_x$. Thus, $\|w_c^{(M)} - w^*\| \leq \|w_x^{(M)} - w^*\| - \alpha$ for some $\alpha > 0$.

Proof of Claim 2 (Test error)

The test error for a parameter w is $E(w) = |w^\top x^* - y^*|$. Using the given definition $y^* = (w^*)^\top c^* + \varepsilon_y^*$, we can write the error as:

$$E(w) = |(w - w^*)^\top x^* + ((w^*)^\top x^* - ((w^*)^\top c^* + \varepsilon_y^*))|$$

Let $\mathcal{R} := \mathcal{R}(x^*, c^*, w^*) := (w^*)^\top x^* - (w^*)^\top c^* - \varepsilon_y^*$. It captures any discrepancy between the prediction made by w^* using the full x^* and the target y^* defined via c^* . So, $E(w) = |(w - w^*)^\top x^* + \mathcal{R}(x^*, c^*, w^*)|$.

Let $\Delta w_c = w_c^{(M)} - w^*$ and $\Delta w_x = w_x^{(M)} - w^*$. From Claim 1, we know that $\|\Delta w_c\| \leq \|\Delta w_x\| - \alpha$. The respective test errors are as follows.

$$E(w_c^{(M)}) = |(\Delta w_c)^\top x^* + \mathcal{R}| \quad \text{and} \quad E(w_x^{(M)}) = |(\Delta w_x)^\top x^* + \mathcal{R}|.$$

By the Cauchy-Schwarz inequality,

$$|(\Delta w_c)^\top x^*| \leq \|\Delta w_c\| \|x^*\| \leq (\|\Delta w_x\| - \alpha) \|x^*\|.$$

By the triangle inequality applied to $E(\cdot)$,

$$\begin{aligned} E(w_x^{(M)}) &\geq |(\Delta w_x)^\top x^*| - |\mathcal{R}|, \\ E(w_c^{(M)}) &\leq |(\Delta w_c)^\top x^*| + |\mathcal{R}|. \end{aligned}$$

Hence,

$$E(w_x^{(M)}) - E(w_c^{(M)}) \geq (|(\Delta w_x)^\top x^*| - |(\Delta w_c)^\top x^*|) - 2|\mathcal{R}|.$$

Under Assumption C.5,

$$|(\Delta w_x)^\top x^*| - |(\Delta w_c)^\top x^*| \geq \rho \|x^*\| \alpha.$$

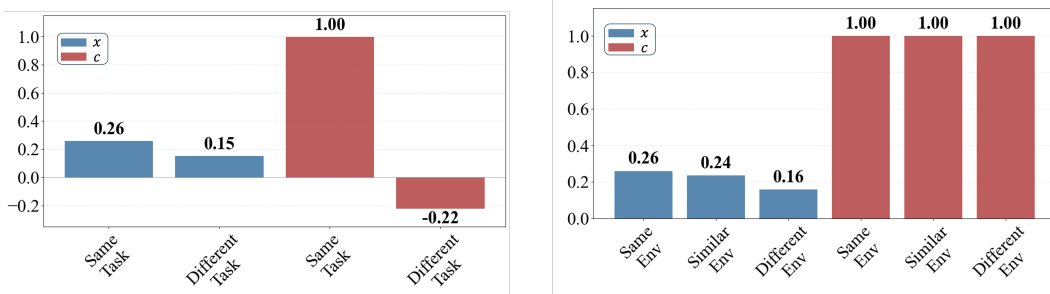
Then the difference of error functions has the following lower bound,

$$E(w_x^{(M)}) - E(w_c^{(M)}) \geq \underbrace{\rho \|x^*\| \alpha - 2|\mathcal{R}|}_{=: \alpha'}.$$

Equivalently,

$$E(w_c^{(M)}) \leq E(w_x^{(M)}) - \alpha'.$$

Under Assumption C.6, $\alpha' > 0$.



(a) Cosine similarities between two datasets with the same environments

(b) Cosine similarities between two datasets with the same tasks

Figure 2: Cosine similarity comparisons for two representations, x (blue), and the learned causal representation, c (red), under different task or environment conditions. In (a), we fix the same environment and vary the task type (e.g., “Same Task”, or “Different Task”). In (b), we fix the same task and vary the environment (e.g., “Same Env”, “Similar Env”, or “Different Env”). The raw x -based similarities remain relatively small or inconsistent across both task and environment variations, whereas the c -based similarities clearly distinguish between task-relevant and environment-relevant differences. This result indicates that c focuses on capturing task-level invariance while filtering out environment-specific noise.

D Detailed experiment settings

D.1 Synthetic dataset

We conduct a toy experiment in order to verify (1) whether CCL learns domain-invariant (causal) features and (2) how much CCL can distinguish samples rigorously at each task and environment pair.

To this end, we consider three distinct tasks across five different environments. To model the data-generating process, we represent each task t and each domain-specific factor s using 64-dimensional random embeddings. These embeddings are constructed to be mutually orthogonal, ensuring disentangled representations of task- and environment-specific variations. After that, a simple two-layer MLP with random Gaussian noise is used to generate the c variable. To disambiguate between tasks, we then conduct contrastive learning to ensure that c embeddings generated from the same task are similar and otherwise distinct. We follow a similar process when generating the remaining variables x , y , and e . In this case, we categorize the relationships between environments into three types: ‘Same’, ‘Different’, and ‘Similar’. When contrastive learning on two of the five environments that are tuned to be similar, set the weight to 0.5 to regulate them from being too different. The x variable follows the same process as generating y or e , except that it is concatenated with c and s as input to the MLP.

Figure 2 compares the embedding of overall samples in a single environment or in a single task. Inferred c by CCL shows that sample similarity is task-dependent in the same environment, while within the same task, it is independent of environmental factors. On the other hand, x confirms that sample similarity is affected by both task and environment. This shows that CCL can estimate the domain-invariant variable c well.

D.2 MGSM

D.2.1 ID / OOD dataset split

We classify the MGSM dataset into ID and OOD using the x embeddings obtained from the text-embedding-3-small model. To accomplish this, we extract x embeddings for each question and then retrieve the closest samples using KNN without distinguishing between ID and OOD data. Figure 3 presents a heatmap illustrating which language’s questions are retrieved as the closest samples for each language. In Figure 3, each row represents the input question’s language. The heatmap for each row shows, among 250 questions in that language, which languages appear as the closest samples. For instance, for 250 English questions, when retrieving the top-1 closest samples: 45 questions are

matched with German samples, 95 questions are matched with French samples, 109 questions are matched with Spanish samples.

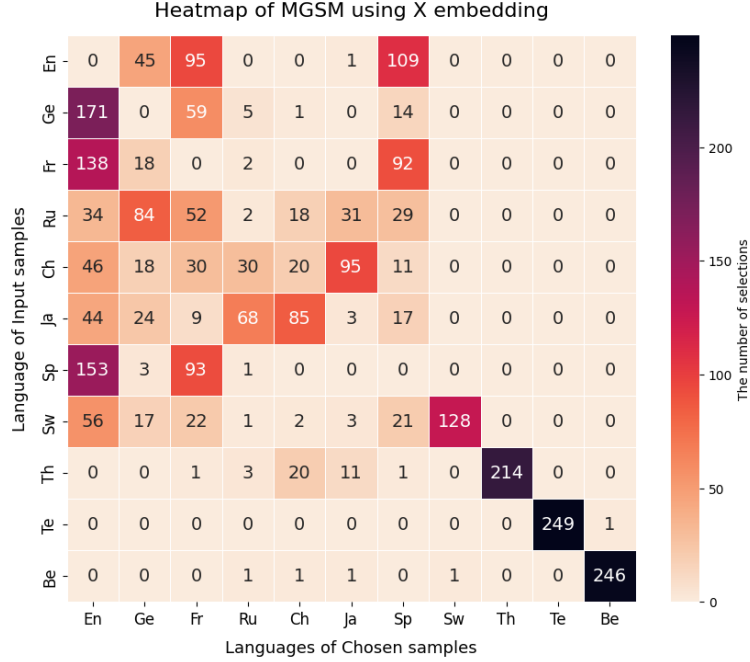


Figure 3: The heatmap of the MGSM dataset using x embedding

Using this heatmap, we categorize the dataset into ID and OOD. We analyze the four languages defined as OOD: Swahili (Sw), Thai (Th), Telugu (Te), and Bengali (Be). For the remaining languages (excluding these four), we observe that they frequently appear as the retrieval output language for various input languages in the top-1 closest sample. For example, English is retrieved as the closest sample for 171 German questions. However, the four OOD languages are rarely selected as the closest retrieved sample for languages other than themselves. Furthermore, in the MGSM dataset, the same problem does not exist within the same language. Therefore, retrieving samples from the same language is not helpful for problem-solving. Nevertheless, unlike other languages, the OOD languages tend to retrieve samples from their own language with significantly higher probability.

Considering these factors, we classify Swahili (Sw), Thai (Th), Telugu (Te), and Bengali (Be) as the OOD dataset in this experiment.

D.2.2 Task specific prompt

The prompts for the MGSM dataset are based on OpenAI’s Simple-evals².

Role	Prompts
System	"You are a helpful assistant."
User	Solve this math problem. Give the reasoning steps before giving the final answer on the last line by itself in the format of "Answer:". Do not add anything other than the integer answer after "Answer:". Question: {Sample 1 Question}. Answer: {Sample 1 Answer} Question: {Sample 2 Question}. Answer: {Sample 2 Answer} ... Question: {input}. Answer:

Table 1: System and User Prompt utilized in MGSM dataset

²<https://github.com/openai/simple-evals>

Table 1 shows the prompts utilized in MGSM datasets. The same prompts are used for all methods except Zero-shot, and Sample Question and Sample Answer contain questions and answers from in-context samples. The input contains the actual question for the model to solve, and the model reports its answer after “Answer: ”.

Role	Prompts
System	"Do not repeat the given input. Respond concisely with a single-word or short-phrase answer."
User	<p>### Instruction ###</p> <p>The following information provides important context to help generate an appropriate answer. Analyze it carefully before responding.</p> <p>### Example ###</p> <p>{sample 1 Question}</p> <p>{sample 1 Answer}</p> <p>{sample 2 Question}</p> <p>{sample 2 Answer}</p> <p>...</p> <p>{sample 8 Question}</p> <p>{sample 8 Answer}</p> <p>### Input ###</p> <p>{Input Query}</p> <p></Answer>positive/negative</Answer></p>

Table 2: System and User Prompt utilized in OOD NLP tasks

Table 2 shows the prompts utilized in the sentiment analysis task in the OOD NLP experiment. We set the options for each task.

D.2.3 Fixed in-context samples

For ICL (Fix.), we obtain the fixed samples from Google Research’s `Url-nlp`³. `Url-nlp` provides the same 8 example problems translated into 10 languages. We randomly select five samples for each language for the 5-shot experiment, and use the same problems for all languages. Table 3 shows the questions and answers of the samples utilized in our experiments, representing only the samples written in English.

Question	Answer
Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?	11
Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	39
Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?	33
Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?	8
Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?	8

Table 3: Fixed 5 samples randomly obtained from `Url-nlp`

D.3 Advanced experiments

The unseen generation task focuses on *sentiment reversal paraphrasing*, where the model rewrites a negative sentence to express an opposite (positive) sentiment. We automatically evaluate the

³<https://github.com/google-research/url-nlp>

generated outputs with *GPT-4o-mini*, which performs sentiment analysis to verify whether the reversal is successful. Although CCL trains only on binary and multiple-choice classification tasks, it generalizes well to this unseen generation setting.

We next evaluate CCL on the MMLU benchmark to test its reasoning ability. We follow the common 5-shot setting using the "lukaemon/mmlu" dataset, which provides QA pairs across 57 diverse domains. We use samples from the training split as demonstrations and the validation split for evaluation. We combine both splits to train CCL and build a VAE across all 57 domains. For each query, CCL retrieves five embedding-level examples across domains, leveraging its ability to recognize query intent and select semantically relevant examples. We also include ICL results that use the same embedding-based retrieval strategy.

We further test CCL on HotpotQA, a multi-hop QA benchmark that requires integrating evidence through stepwise reasoning. We provide each target query with a relevant document, since CCL is not a retriever like RAG. We further treat each example collected at the *c*-level or *x*-level, together with its corresponding document, as a single shot. This setup investigates whether intent-aligned examples enable the model to learn a more effective reasoning process for deriving answers from a given document, compared with examples selected solely on surface-level similarity.

References

- [1] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.