Context-Adaptive Inference: A Unified Statistical and Foundation-Model View

Authors

- **Yue Yao**
- **Caleb N. Ellington**
- **Jingyun Jia**
- **Baiheng Chen**
- **Dong Liu**
- **Rikhil Rao**
- **Jiaqi Wang**
- **Samuel Wales-McGrath**
- **Yixin Yang**
- **Zhiyuan Li**
- **Ben Lengerich** –

Abstract

context-adaptive inference

c $\theta(c)$ $f(x; \theta(c))$

Introduction

context→parameters

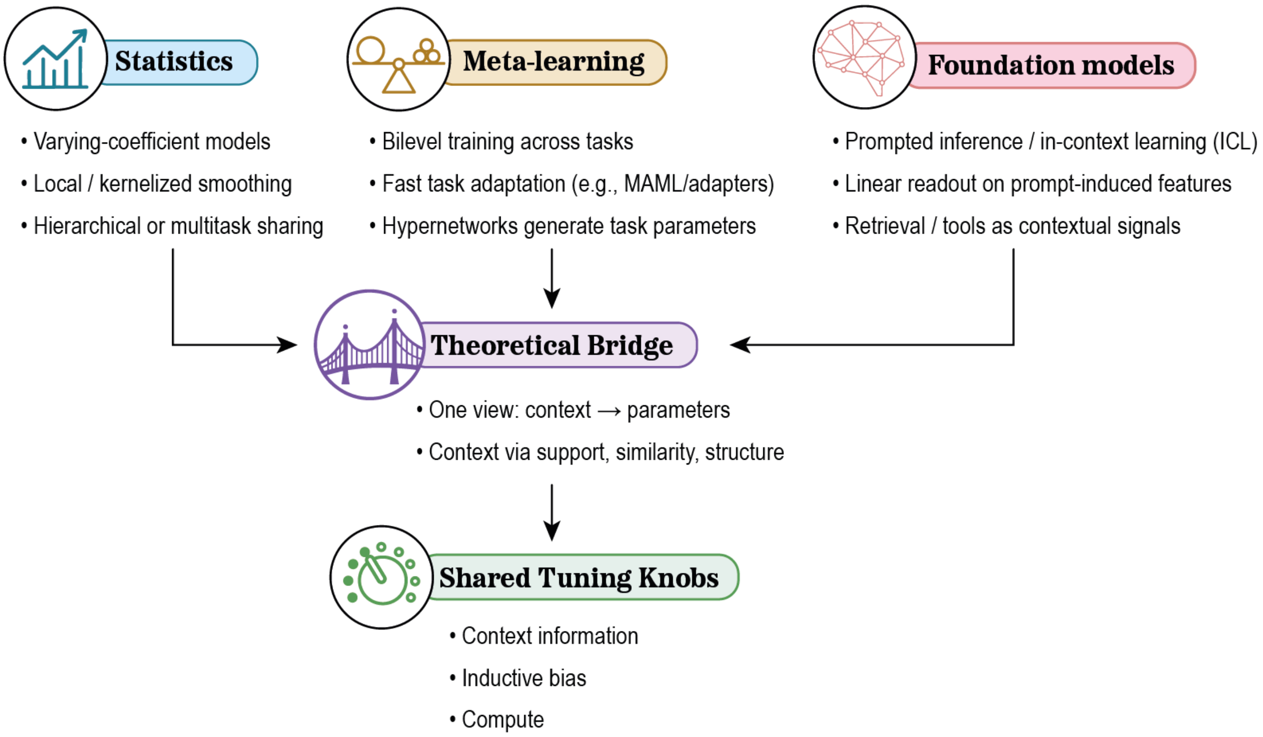

Statistics
• Varying-coefficient models
• Local / kernelized smoothing
• Hierarchical or multitask sharing
Meta-learning
• Bilevel training across tasks
• Fast task adaptation (e.g., MAML/adapters)
• Hypernetworks generate task parameters
Foundation models
• Prompted inference / in-context learning (ICL)
• Linear readout on prompt-induced features
• Retrieval / tools as contextual signals
Theoretical Bridge
• One view: context ⟶ parameters
• Context via support, similarity, structure
Shared Tuning Knobs
• Context information
• Inductive bias
• Compute

Figure 1:

x_i

θ_i

$$x_i \sim P(x;\, \theta_i).$$

$\theta_i = \theta$ i

$$\theta_i = f(c_i) \quad \text{or} \quad \theta_i \sim P(\theta \mid c_i),$$

c_i i

f

θ_i

Problem Setup and Notation

$i = 1, \ldots, n$ **context** $c_i \in \mathcal{C}$

$\mathcal{D}_i = \{(x_{ij}, y_{ij})\}_{j=1}^{m_i}$ $x_{ij} \in \mathcal{X}$ $y_{ij} \in \mathcal{Y}$

$\mathcal{H} = \{h_\theta : \mathcal{X} \to \mathcal{Y} \mid \theta \in \Theta\}$

global $\theta_i \equiv \theta^\star$ **context-adaptive**

$\theta_i = f(c_i)$ $\theta_i \sim P(\theta \mid c_i)$

c

$$\hat{\theta}(c) \in \arg\min_{\theta\in\Theta} \underbrace{\sum_{(i,j)\in S(c)} \ell\big(h_\theta(x_{ij}), y_{ij}\big)}_{\text{context-dependent support}} + \underbrace{\mathcal{R}(\theta;\, c)}_{\text{context-structured regularization}} , \qquad (\star)$$

ℓ $S(c) \subseteq \{1,\dots,n\} \times \mathbb{N}$ **support set**
c $\mathcal{R}(\theta; c)$

How context enters.
Explicit parameterization: $f : \mathcal{C} \to \Theta$ $\theta_i = f(c_i)$
$\mathcal{R}(\theta; c)$ f $\mathcal{C}$
Implicit parameterization:
θ $g(x, c)$ $S(c)$
$R(c)$ $P(c)$

context encoder $\phi : \mathcal{C} \to \mathbb{R}^d$ $K(c, c')$

$$\sum_{i,j} w_{ij}(c)\, \ell\big(h_\theta(x_{ij}), y_{ij}\big) \;+\; \mathcal{R}(\theta), \qquad w_{ij}(c) \propto K\big(\phi(c), \phi(c_i)\big) \cdot \mathbf{1}\big[(i,j) \in S(c)\big].$$

Granularity. $g \in \{\text{group}, \text{unit}, \text{example}\}$

Information $S(c)$ $P(c)$
Inductive bias $\mathcal{R}(\theta; c)$
Compute

Standing assumptions (used as needed).
(θ_i, c_i) (x_{ij}, y_{ij})
$\theta = f(c)$ f
$w_{ij}(c)$
ℓ $\mathcal{R}$
$|S(c)|$ **adaptation efficiency**

Theoretical Bridge

Proposition 1 (Explicit varying-coefficients and linear ICL coincide with kernel ridge on joint features in the linear squared-loss setting).
$y = \langle \theta(c), x \rangle + \varepsilon$ $\mathbb{E}[\varepsilon] = 0$
$\phi : \mathcal{C} \to \mathbb{R}^{d_c}$ $\psi(x, c) := x \otimes \phi(c) \in \mathbb{R}^{d_x d_c}$
$S(c)$ $w_{ij}(c)$

- **(A) Explicit varying-coefficients.** $\theta(c) = B\,\phi(c)$ $B \in \mathbb{R}^{d_x \times d_c}$ $\lambda\|B\|_F^2$

$$\hat{y}(x,c) = k_{(x,c)}^{\top}\big(K + \lambda I\big)^{-1} y, \quad K_{ab} = \langle \psi_a, \psi_b \rangle = \langle x_a, x_b \rangle \cdot \langle \phi(c_a), \phi(c_b) \rangle,$$

kernel ridge regression (KRR)

- **(B) Implicit adaptation via linear ICL.** $S(c)$ $q = Q\psi$ $k = K\psi$ $v = V\psi$ $w_{ij}(c) \cdot \langle q, k_{ij} \rangle$

$$k\big((x,c),(x',c')\big) = \langle q(x,c), k(x',c') \rangle,$$

dot-product kernel

ψ

Corollary 1 (Retrieval, gating, and weighting are kernel/measure choices). $S(c)$ $R(c)$ $w_{ij}(c)$ ψ

context-aware $S(c)$ $\mathcal{R}(\theta; c)$

Scope of Review and Relation to Prior Work

Related Surveys and Reviews

Survey	Topic Focus	Scope	Coverage of Adaptivity	Gap Relative to This Work

Survey	Topic Focus	Scope	Coverage of Adaptivity	Gap Relative to This Work

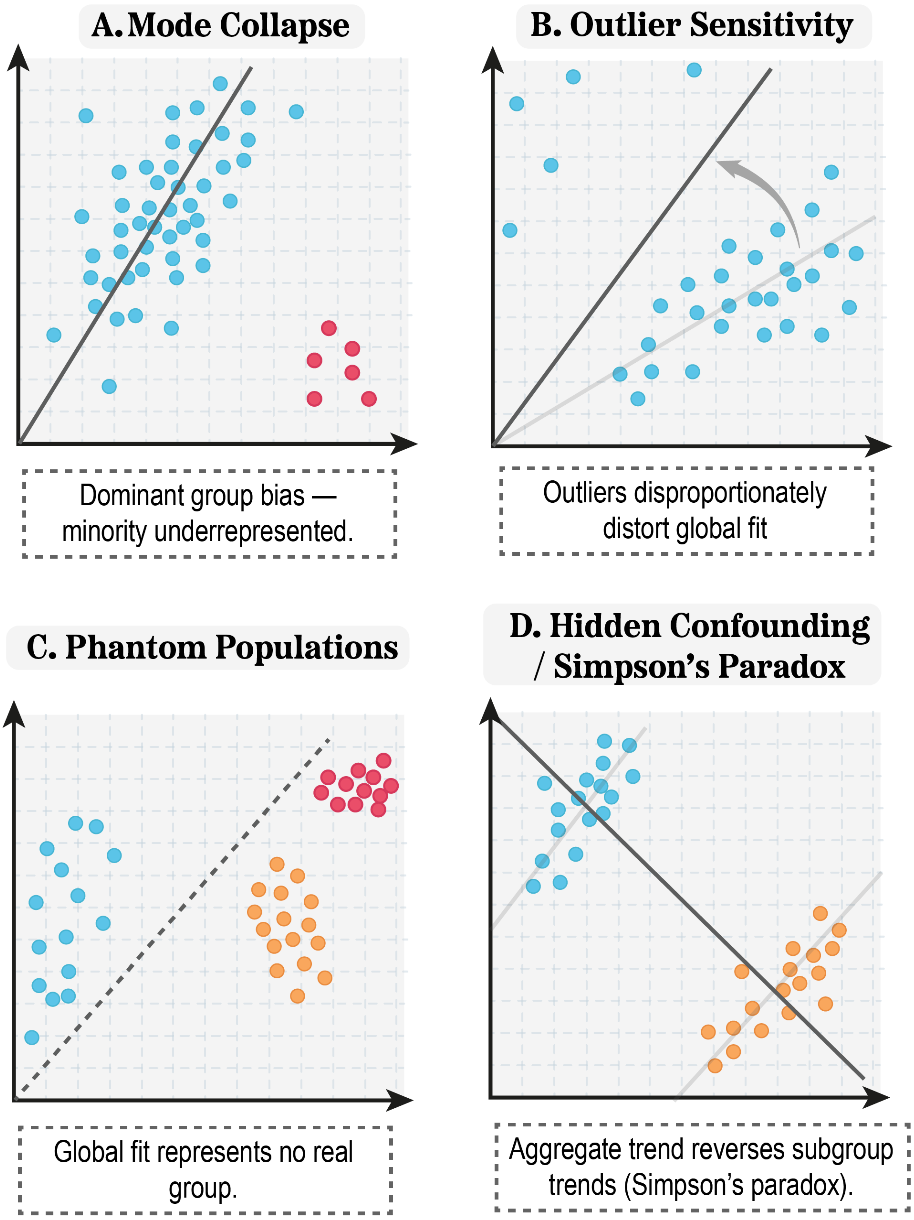
From Population Assumptions to Context-Adaptive Inference

Failure Modes of Population Models

Mode Collapse

Outlier Sensitivity

Phantom Populations



A. Mode Collapse

Dominant group bias — minority underrepresented.

B. Outlier Sensitivity

Outliers disproportionately distort global fit

C. Phantom Populations

Global fit represents no real group.

D. Hidden Confounding / Simpson's Paradox

Aggregate trend reverses subgroup trends (Simpson's paradox).

Figure 2:

Mode Collapse

Outlier Sensitivity

Phantom Populations **Hidden Confounding**

/ Simpson's Paradox

Toward Context-Aware Models

$$x_i \sim P(x;\theta_i),$$

N N θ_i c_i

$$\theta_i = f(c_i) \quad \text{or} \quad \theta_i \sim P(\theta \mid c_i).$$

f

$Y(1)$ $Y(0)$ $E[Y(1) - Y(0)]$

X C

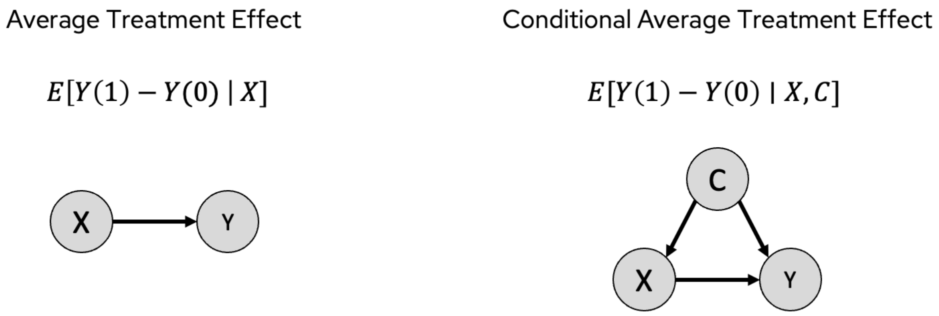

Average Treatment Effect
$E[Y(1) - Y(0) \mid X]$
X
Y
Conditional Average Treatment Effect
$E[Y(1) - Y(0) \mid X, C]$
C
X
Y

Figure 3: X C

$f(c)$

Classical Remedies: Grouped and Distance-Based Models

$f(c)$

Conditional and Clustered Models

$$\{\hat{\theta}_0, \ldots, \hat{\theta}_C\} = \arg \max_{\theta_0, \ldots, \theta_C} \sum_{c \in \mathcal{C}} \ell(X_c; \theta_c),$$

$\ell(X;\theta)$ θ X c

Distance-Regularized Estimation

θ_i

c_i

$$\{\hat{\theta}_0, \ldots, \hat{\theta}_N\} = \arg \max_{\theta_0, \ldots, \theta_N} \left(\sum_i \ell(x_i; \theta_i) - \sum_{i,j} \frac{\|\theta_i - \theta_j\|}{D(c_i, c_j)} \right),$$

$D(c_i, c_j)$

D λ

Parametric and Semi-parametric Varying-Coefficient Models

$$\widehat{A} = \arg\max_A \sum_i \ell(x_i; Ac_i).$$

Contextualized Models

$f(c)$ f f

$$\widehat{f} = \arg\max_{f \in \mathcal{F}} \sum_i \ell(x_i; f(c_i)).$$

Partition and Latent-Structure Models

$$\{\hat{\theta}_0, \ldots, \hat{\theta}_N\} = \arg\max_{\theta_0, \ldots, \theta_N} \left(\sum_i \ell(x_i; \theta_i) + \lambda \sum_{i=2}^{N} \|\theta_i - \theta_{i-1}\| \right).$$

Fine-tuned Models and Transfer Learning

Models for Explicit Subgroup Separation

A Spectrum of Context-Awareness

- **Global models** $\theta_i = \theta$ i
- **Grouped models** $\theta_i = \theta_c$
- **Smooth models** $\theta_i = f(c_i)$ f
- **Latent models** $\theta_i \sim P(\theta \mid c_i)$ f

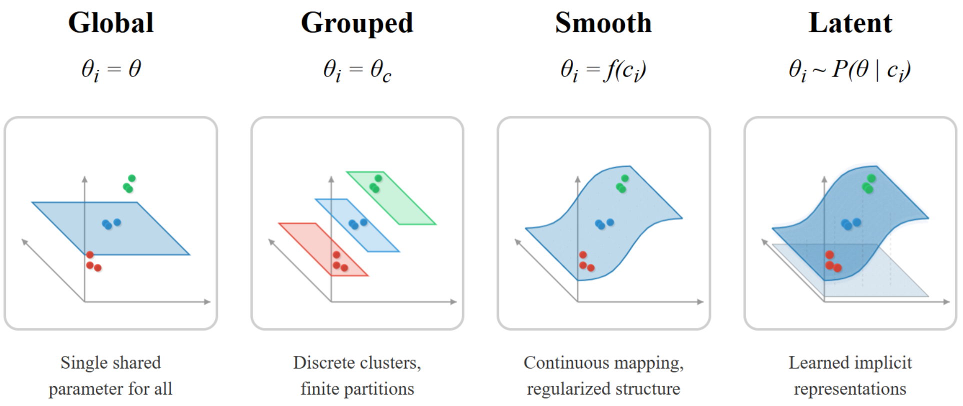

Global
$\theta_i = \theta$
Single shared parameter for all
Grouped
$\theta_i = \theta_c$
Discrete clusters, finite partitions
Smooth
$\theta_i = f(c_i)$
Continuous mapping, regularized structure
Latent
$\theta_i \sim P(\theta \mid c_i)$
Learned implicit representations

Figure 4:

Independent and identically distributed samples

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2$$

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

$$g(\mu_i) = \eta_i$$

$$\log \frac{p_i}{1 - p_i} = x_i^\top \beta$$

$$\log(\mu_i) = x_i^\top \beta$$

$$F = \frac{MS_{\text{Between}}}{MS_{\text{Within}}}$$

Hierarchical data

$$y_{ij} = \mu + u_j + \varepsilon_{ij}, \quad u_j \sim N(0, \sigma_u^2), \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$y = X\beta + Zu + \varepsilon, \quad u \sim N(0, G), \quad \varepsilon \sim N(0, R)$$

$$g(\mu_i) = x_i^\top \beta + z_i^\top u$$

β u

$$y_{ij} \mid \theta_j \sim p(y_{ij} \mid \theta_j), \quad \theta_j \sim p(\theta_j \mid \phi), \quad \phi \sim p(\phi)$$

Functional types and high-dimensional data

$x_i(t)$

$x_i(t)$

$$y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i$$

f_j

$h_\phi(x)$ $g_\theta(z)$

$$(\theta^{ast}, \phi^{ast}) = \arg\min_{\theta,\phi} \sum_{i=1}^{n} |x_i - g_\theta(h_\phi(x_i))|^2$$

$z_i = h_\phi(x_i)$ x_i

Heterogeneous tasks and sparse data

T (X^t, Y^t) $t = 1, \ldots, T$ w^t

$$\min_W \sum_{t=1}^{T} \sum_{i=1}^{n_t} \ell(y_i^t, f(x_i^t; w^t)) + \lambda\, \Omega(W)$$

$W = [w^1, \ldots, w^T]$ $\Omega(W)$

$p(x)$ $p(y|x)$

$p_{\text{train}}(x)$ $p_{\text{test}}(x)$

$$\mathbb{E}_{x\sim p_{\text{test}}}[\ell(f(x), y)] = \mathbb{E}_{x\sim p_{\text{train}}}\left[\frac{p_{\text{test}}(x)}{p_{\text{train}}(x)}\, \ell(f(x), y)\right], \quad w(x) = \frac{p_{\text{test}}(x)}{p_{\text{train}}(x)}$$

$p(x)$ $p(y|x)$

$\phi(x)$

Online and interactive data

$t = 1, \ldots, T$

$x_t \in F$ $F \subset \mathbb{R}^n$

$c_t : F \to \mathbb{R}$ $c_t(x_t)$

$$R(T) = \sum_{t=1}^{T} c_t(x_t) - \min_{x \in F} \sum_{t=1}^{T} c_t(x)$$

$g_t \in \partial c_t(x_t)$

$$x_{t+1} = \Pi_F(x_t - \eta_t g_t)$$

η_t Π_F F

$R(T) = O(\sqrt{T})$

x y

$\hat{p}\_i$ i s_i

$(p\_\text{min}, s\_\text{min})$

$$\hat{p}_i + s_i \geq p_{\min} + \alpha \cdot s_{\min}$$

α

k_w k_d $(p_{\min}, s_{\min})$

$c_t(\cdot)$

ϵ $1-\epsilon$

ϵ

t a

$$\mathrm{UCB}_a(t) = \hat{\mu}_a + \sqrt{\frac{2\ln t}{n_a}}$$

$\arg\max_a \mathrm{UCB}_a(t)$

t x_t $\pi : X \to A$

t

$$a_t = \arg\max_{a\in\mathcal{A}} \left(x_t^\top \hat{\theta}_a + \alpha\sqrt{x_t^\top A_a^{-1} x_t} \right)$$

$\hat{\theta}_a$ A_a

S A

$P(s' \mid s, a)$ $r(s,a)$ $\gamma \in [0,1)$ t

s_t $a_t \sim \pi(\cdot \mid s_t)$ r_t

s_{t+1} π

$$J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

Multimodal data

$$f_\theta : \mathcal{X} \to \mathcal{Z}$$

$\mathcal{X}$ $\mathcal{Z}$

$p(z|x)$

$p_\theta(x|z)$

$q(z|x)$

$q_\phi(z|x)$

$p_\theta(x|z)$ $q_\phi(z|x)$

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x) \,||\, p(z)]$$

$q_\phi(z|x)$

$$\min_\theta \sum_{T_i \sim p(T)} \mathcal{L}_{T_i}\big(U(\theta, T_i)\big)$$

θ T_i

$p(T)$ $U(\theta, T_i)$

θ

Large-scale pre-trained data

$\mathcal{D} = \{x_i\}_{i=1}^{N}$

f_θ

$$\theta^* = \arg\min_{\theta} \ \mathbb{E}_{x \sim \mathcal{D}} \, \ell(f_\theta(x))$$

ℓ

$\{(x_i, y_i)\}_{i=1}^{k}$ x_{k+1}

$\hat{y}_{k+1}$

$$\hat{y}_{k+1} = f_\theta(x_{k+1} \mid x_1, y_1, \ldots, x_k, y_k)$$

Principles of Context-Adaptive Inference

1. Adaptivity requires flexibility

2. Adaptivity requires a signal of heterogeneity

3. Modularity improves adaptivity

4. Adaptivity implies selectivity

5. Adaptivity is bounded by data efficiency

6. Adaptivity is not a free lunch

When Adaptivity Fails: Common Failure Modes

Spurious adaptation.

Overfitting in low-data contexts.

Modularity mis-specification.

Feedback loops.

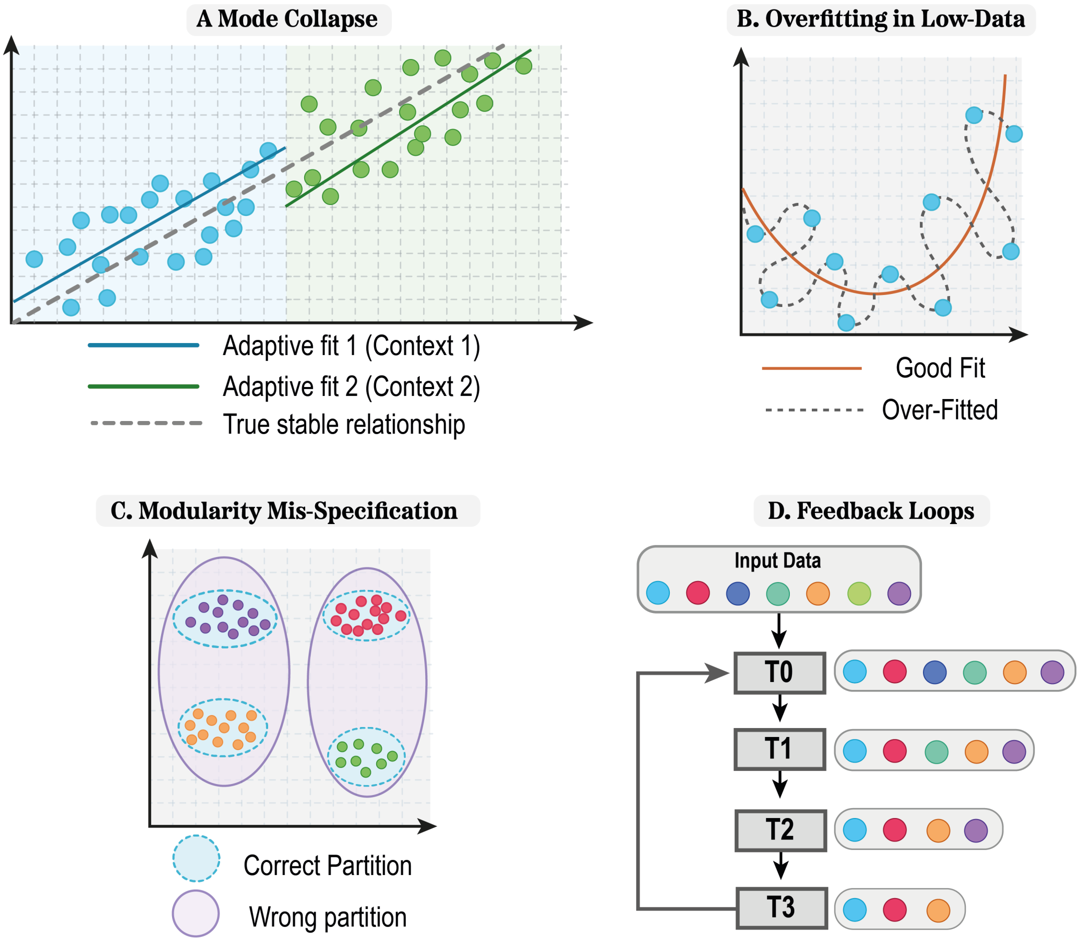


Figure 5: **Mode Collapse** **Overfitting in Low-Data Contexts** **Modularity Mis-Specification** **Feedback Loops**

Synthesis and Implications

Context-Aware Efficiency Principles and Design

Adaptivity is bounded by data efficiency

Formalization: data-efficiency constraints on adaptivity

$\mathcal{C}$

$\theta(c) \in \Theta$

(x, y, c) $p_\theta(y \mid x, c)$ $\ell(\theta; x, y, c)$

$\mathcal{N}_\delta(c) = \{c' : d(c, c') \leq \delta\}$ d

$\theta(c)$

$$N_{\text{eff}}(c, \delta) = \sum_{i=1}^{n} w_\delta(c_i, c), \quad w_\delta(c_i, c) \propto K\left(\frac{d(c_i, c)}{\delta}\right), \quad \sum_i w_\delta(c_i, c) = 1,$$

K

$$\mathcal{R}(\theta) = \int \|\nabla_c \theta(c)\|^2 \, \mathrm{d}c$$

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; x_i, y_i, c_i) + \lambda \mathcal{R}(\theta).$$

c L μ θ

j

$$\mathbb{E}\Big[\|\hat{\theta} j(c) - \theta_j(c)\|^2\Big] \lesssim \underbrace{\frac{\sigma^2}{N_{\text{eff}}(c, \delta)}}_{\text{variance}} + \underbrace{\delta^{2\alpha}}_{\text{approx. bias}} + \underbrace{\lambda^2}_{\text{reg. bias}}, \quad \alpha > 0,$$

δ

N_{eff}

δ λ

$\theta(c)$ $f_\phi(c)$ ϕ N_{eff}

η $T(c)$

$$\mathcal{L}\big(\theta^{(T(c))}\big) - \mathcal{L}(\theta^\star) \;\leq\; (1 - \eta\mu)^{T(c)} \Big(\mathcal{L}\big(\theta^{(0)}\big) - \mathcal{L}(\theta^\star)\Big) + \frac{\eta L \sigma^2}{2\mu \, N_{\text{eff}}(c, \delta)}.$$

$T(c)$ $N_{\text{eff}}(c, \delta)$

c

Formal optimization view of context-aware efficiency

$f_\phi : \mathcal{X} \times \mathcal{C} \to \mathcal{Y}$ ϕ

$T(c)$ $\Omega(\phi)$

$$\min_{\phi} \; \mathbb{E}_{(x,y,c)\sim\mathcal{D}} \ell\big(f_\phi(x, c), y\big) + \lambda \, \Omega(\phi) \quad \text{s.t.} \quad \mathbb{E}_c \, \mathcal{C}\big(f_\phi; T(c), c\big) \leq B,$$

$\mathcal{C}(\cdot)$

$$\min_{\phi}\ \mathbb{E}_{(x,y,c)}\ell\big(f_\phi(x,c),y\big)+\lambda\,\Omega(\phi)+\gamma\,\mathbb{E}_c\,\mathcal{C}\big(f_\phi;T(c),c\big),$$

γ

$\phi=(\phi_1,\dots,\phi_M)$

$\pi_\phi(m\mid c)$

$$\Omega(\phi)=\sum_{m=1}^{M}\alpha_m\,\|\phi_m\|_2^2+\tau\,\mathbb{E}_c\sum_{m=1}^{M}\pi_\phi(m\mid c),$$

$$\nabla_\phi\Big(\mathbb{E}\,\ell+\lambda\,\Omega+\gamma\,\mathbb{E}_c\,\mathcal{C}\Big)=0,\quad \gamma\,\big(\mathbb{E}_c\,\mathcal{C}-B\big)=0,\quad \gamma\geq 0.$$

$T(c)$

Explicit Adaptivity: Structured Estimation of $f(c)$

$\theta_i=f(c_i)$ f

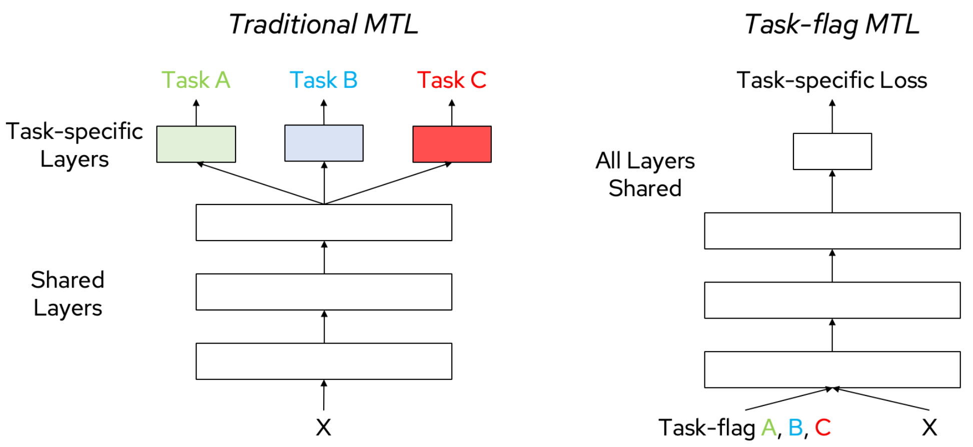


Figure 6:

$f(c)$

$f(c)$

Classical Varying-Coefficient Models: A Foundation

$$y_i = \sum_{j=1}^{p} \beta_j(c_i) x_{ij} + \varepsilon_i$$

$\beta_j(c)$

Advances in Modeling $f(c)$

$f(c)$

Smooth Non-parametric Models

$f(c)$

c

Structured Regularization for Graphical and Network Models

$f(c)$

c

$f(c)$

Piecewise-Constant and Partition-Based Models.

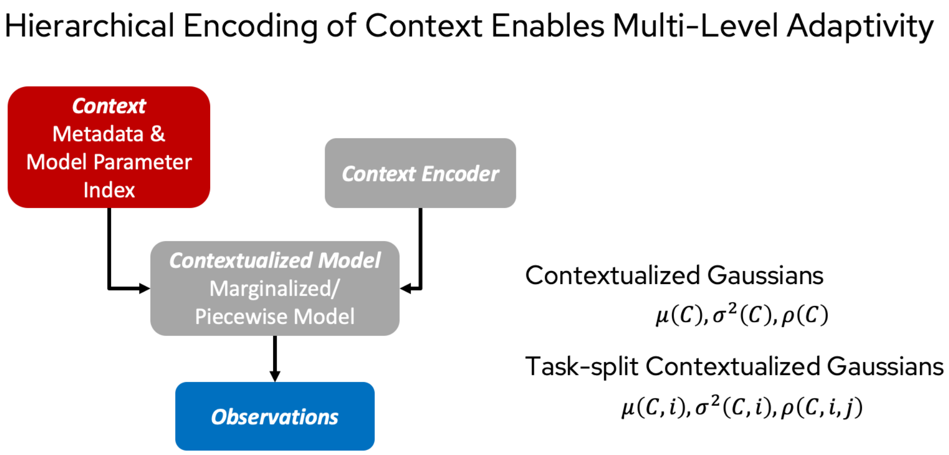

Hierarchical Encoding of Context Enables Multi-Level Adaptivity
Context
Metadata &
Model Parameter
Index
Context Encoder
Contextualized Model
Marginalized/
Piecewise Model
Observations
Contextualized Gaussians
$\mu(C), \sigma^2(C), \rho(C)$
Task-split Contextualized Gaussians
$\mu(C,i), \sigma^2(C,i), \rho(C,i,j)$

Figure 7: c (i,j)

simple parametric models within each context c Z **aggregate across contexts**

$$P(Y \mid X, C) \; = \int P(Y \mid X, C, Z) \, dP(Z \mid C)$$

global flexibility can emerge from compositional, context-specific parametrics

c

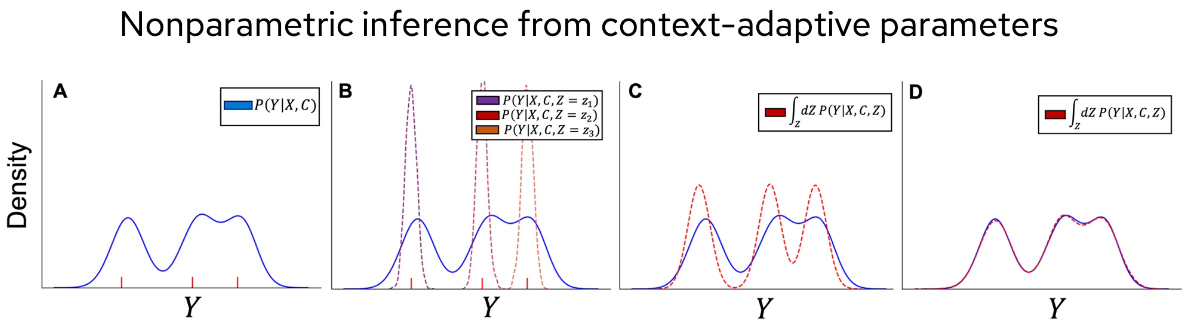


Figure 8: $P(Y \mid X, C)$ $P(Y \mid X, C, Z = z_i)$ Z

$\int_Z P(Y \mid X, C, Z)$

context-to-mixture weights **local parametric maps**

Structured Regularization for Spatial, Graph, and Network Data.

c

Learned Function Approximators

$f(c)$

$f(c)$

Tree-Based Ensembles.

Deep Neural Networks.

$f(c)$

Key Theoretical Advances

Theory for Smooth Non-parametric Models.

Theory for Structurally Constrained Models.

Theory for High-Capacity and Learned Models.

Sparsity and Incomplete Measurements as Context

measurement sparsity itself as context.

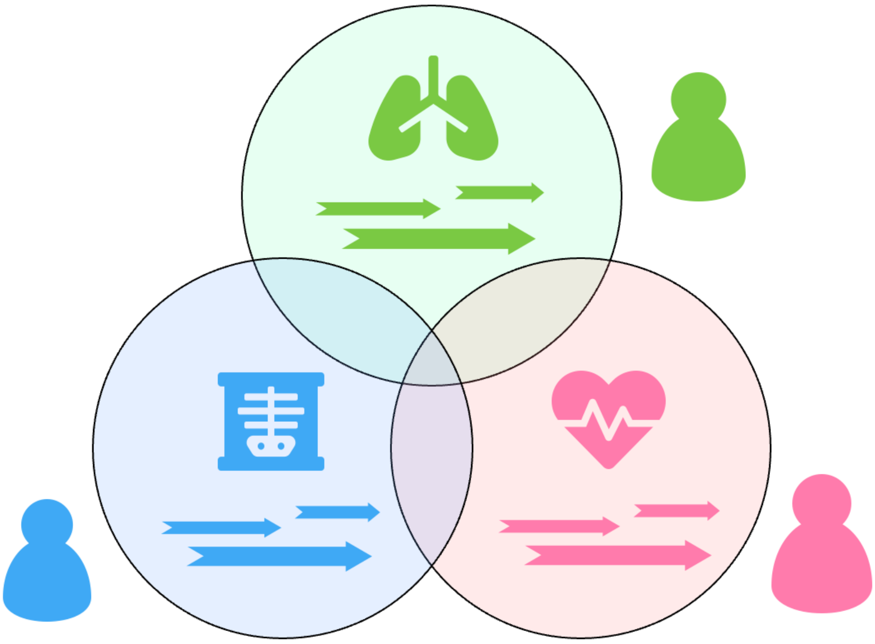


Figure 9:

$p(x_{\text{missing}} \mid x_{\text{observed}})$

GRU-D **BRITS**
GAIN **VAEAC**
XGBoost

Context-Aware Efficiency Principles and Design

Synthesis and Future Directions

$f(c)$

c_i

$f(c)$

Implicit Adaptivity: Emergent Contextualization in Complex Models

Introduction: From Explicit to Implicit Adaptivity.

$\theta_i = f(c_i)$ c_i

θ_i

Foundations of Implicit Adaptation

Architectural Conditioning via Context Inputs

$y_i = g([x_i, c_i]; \Phi)$

c_i x_i g

Φ

x_i y_i c_i

Interaction Effects and Attention Mechanisms

Amortized Inference and Meta-Learning

Amortized Inference

Meta-Learning: Learning to Learn

In-Context Learning in Foundation Models

The Phenomenon of Few-Shot In-Context Learning

Deconstructing ICL: Key Influencing Factors

The Role of Scale.

Prompt Engineering and Example Selection.

Hypothesized Mechanisms: How Does ICL Work?

ICL as Implicit Meta-Learning.

ICL as Implicit Bayesian Inference.

The Role of Induction Heads.

Limitations and Open Questions

Theoretical Bridges Between Varying-Coefficient Models and In-Context Learning

Varying-Coefficient Models as Kernel Regression

θ_i c_i c^*

$$\hat{\theta}(c^*) = \arg\max_{\theta} \sum_{i=1}^{n} K_\lambda(c_i, c^*)\, \ell(x_i; \theta),$$

K_λ ℓ

$y = (y_1, \ldots, y_n)^\top$ $K \in \mathbb{R}^{n \times n}$ $K_{ij} = k(c_i, c_j)$ c^*

$$\hat{y}(c^*) = k(c^*)^\top (K + \lambda I)^{-1} y,$$

$k(c^*) = (k(c^*, c_1), \ldots, k(c^*, c_n))^\top$

c^*

$$\widehat{f}\,(x^*, c^*) = \sum_{i=1}^{n} \alpha_i(c^*)\, y_i,$$

$\alpha_i(c^*)$

λ

Transformers as Ridge and Kernel Regressors In-Context

$$\widehat{w} = (X^\top X + \lambda I)^{-1} X^\top y$$

(x_i, y_i)

x^*

$k(c_i, c_j)$

Synthesis: Two Paths to the Same Estimators

$$\widehat{f}\,(x^*, c^*) = \sum_{i=1}^{n} \alpha_i(c^*)\, y_i,$$

$\alpha_i(c^*)$ c^*
$\{c_i\}$

- $\alpha_i(c^*)$ K_λ
- $\alpha_i(c^*)$

Comparative Synthesis: Implicit versus Explicit Adaptivity

Implicit Adaptivity.

$f(c)$

Explicit Adaptivity. $f(c)$

Open Challenges and the Motivation for Interpretability

Toward Explicit Modeling of Implicit Adaptivity: Local Models, Surrogates and Post Hoc Approximations

Motivation

From Implicit to Explicit Adaptivity

Fidelity vs. Interpretability **Local vs. Global Scope** **Approximation vs. Control**

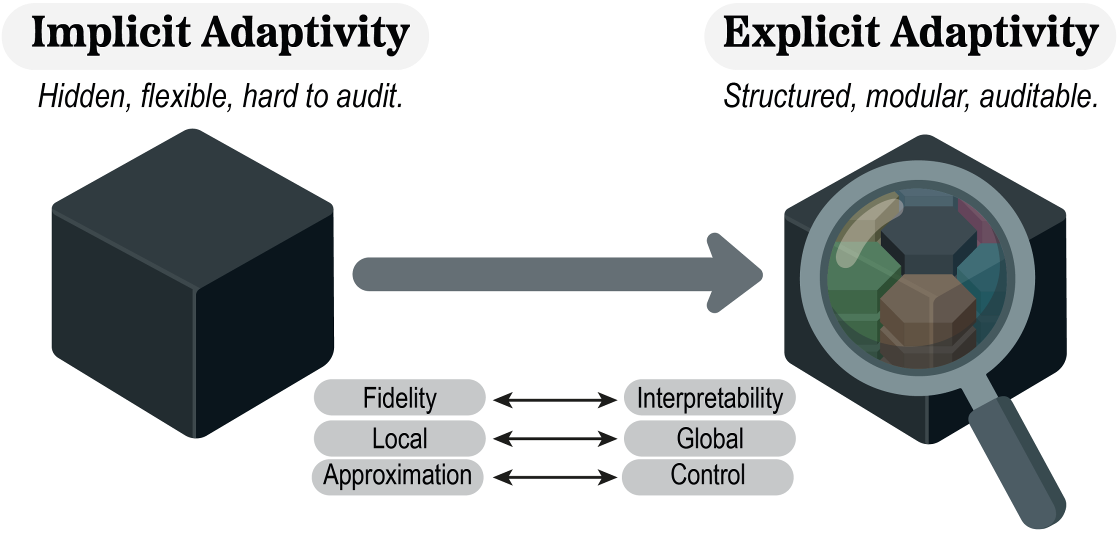


Figure 10:

Approaches

Surrogate Modeling

$h(x, c)$

$f(c)$

$$\hat{g}_{x_0,c_0} = \arg\min_{g \in \mathcal{G}} \mathbb{E}_{(x,c)\sim\mathcal{N}_{x_0,c_0}} \left[\ell\big(h(x,c), g(x,c)\big)\right] + \Omega(g),$$

$\mathcal{N}_{x_0,c_0}$ ℓ Ω

$\mathcal{G}$

$$R^2_{\text{local}} = 1 - \frac{\sum_i w_i \left(h_i - g_i\right)^2}{\sum_i w_i \left(h_i - \bar{h}\right)^2}, \qquad w_i \propto \kappa\big((x_i, c_i), (x_0, c_0)\big).$$

Prototype and Nearest-Neighbor Methods

k

Amortization Diagnostics

$q_\phi(\theta \mid x)$ $f(c)$

Disentangled and Bottlenecked Representations

β

Parameter Extraction and Probing

LLMs as Post-hoc Explainers

Trade-offs

Fidelity vs. Interpretability

$$\min_{g\in\mathcal{G}} \underbrace{\phi_{\text{fid}}(g;U)}_{\text{faithfulness on use set }U} + \lambda$$

$$underbrace\psi_{\text{simplicity}}(g)_{\text{sparsity / size / semantic load}},$$

ϕ_{fid} R^2 h $\psi_{\text{simplicity}}$

Local vs. Global Scope

$g_{x_0,c_0} \approx h$ $\mathcal{N}_{x_0,c_0}$ $g_{\text{global}} \approx h$

$$g(x,c) = \sum_{k=1}^{K} w_k(x,c)\, g_k(x,c), \qquad \sum_k w_k(x,c) = 1, \quad w_k \geq 0,$$

g_k w_k

Approximation vs. Control

Open Research Directions

Reusable Modules

Performance Gains

Abstraction Level

Evaluation and Reporting Standards for Classical Post-hoc Methods

Scope and locality

Attribution methods in practice

Faithfulness and robustness

R^2

$\widetilde{\mathrm{AUC}}_S$ $\widetilde{F}_{1,S}$

Minimal reporting checklist

Item	Description
Data slice and context definition	
Surrogate specification and regularization details	
Faithfulness and robustness metrics	R^2
Sensitivity and uncertainty analysis	
Computational constraints	
Observed limitations and failure modes	

Table 2. Minimal Reporting Checklist for Post-hoc Explanations

From post hoc analysis to design

Implications for classical models

Context-Invariant Training: A View from the Converse

Φ w

$R^e(\cdot)$ e

$w = 1$

$P(X)$ $P(Y|X)$

(Y, ID)

Adversarial Robustness as Context-Invariant Training

$x' = x + \delta$ $\|\delta\|_p \leq \varepsilon$

Perception Robustness

Training methods for Context-Invariant Models

$$\min_{f} \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i)$$

$$\min_{f} \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim P_g} [L(f(x), y)]$$

$\mathcal{G}$ P_g

g

strong regularization

Applications, Case Studies, Evaluation Metrics, and Tools

Implementation Across Sectors

Context-Aware Efficiency in Practice

Formal Metrics for Evaluating Context-Aware Performance

$\mathcal{C}$ $\mathcal{D}_{\text{test}}$ (x, y, c)

$\hat{f}$

$$\mathcal{R}(\hat{f} \mid c) = \mathbb{E}\Big[\,\ell\big(\hat{f}(x,c), y\big) \,\Big|\, c\Big], \qquad \mathcal{R}(\hat{f}) = \mathbb{E}_{c\sim\mathcal{D}_{\text{test}}}\Big[\mathcal{R}(\hat{f} \mid c)\Big].$$

$\mathcal{R}(\hat{f} \mid c)$

$\int \mathcal{R}(\hat{f} \mid c)\, \mathrm{d}\Pi(c)$ Π

Adaptation Efficiency

$S_k(c) = \{(x_j, y_j, c)\}_{j=1}^{k}$ k c

$$\mathrm{AE}_k(c) = \mathcal{R}(\hat{f}_0 \mid c) - \mathcal{R}(\hat{f}_{S_k} \mid c), \qquad \mathrm{AE}_k = \mathbb{E}_c\big[\,\mathrm{AE}_k(c)\,\big],$$

$\hat{f}_0$ $\hat{f}_{S_k}$

$k \mapsto \mathrm{AE}_k$

Transfer Performance

$\mathcal{C}_{\mathrm{src}} \to \mathcal{C}_{\mathrm{tgt}}$

ϕ

$$\mathrm{TP}(\phi) = \mathcal{R}_{\mathcal{C}_{\mathrm{tgt}}}(\hat{f}_{\phi}) - \mathcal{R}_{\mathcal{C}_{\mathrm{tgt}}}(\hat{f}_{\mathrm{scratch}}),$$

ϕ

Robustness to Context Shift

Q f

$$\mathrm{RS}(\hat{f}\,; Q) = \sup_{\widetilde{\mathcal{D}} \in Q} \left[\mathcal{R}_{\widetilde{\mathcal{D}}}(\hat{f}\,) - \mathcal{R}_{\mathcal{D}_{\mathrm{test}}}(\hat{f}\,) \right],$$

Context-Aware Efficiency in Practice

Contextualized Network Inference

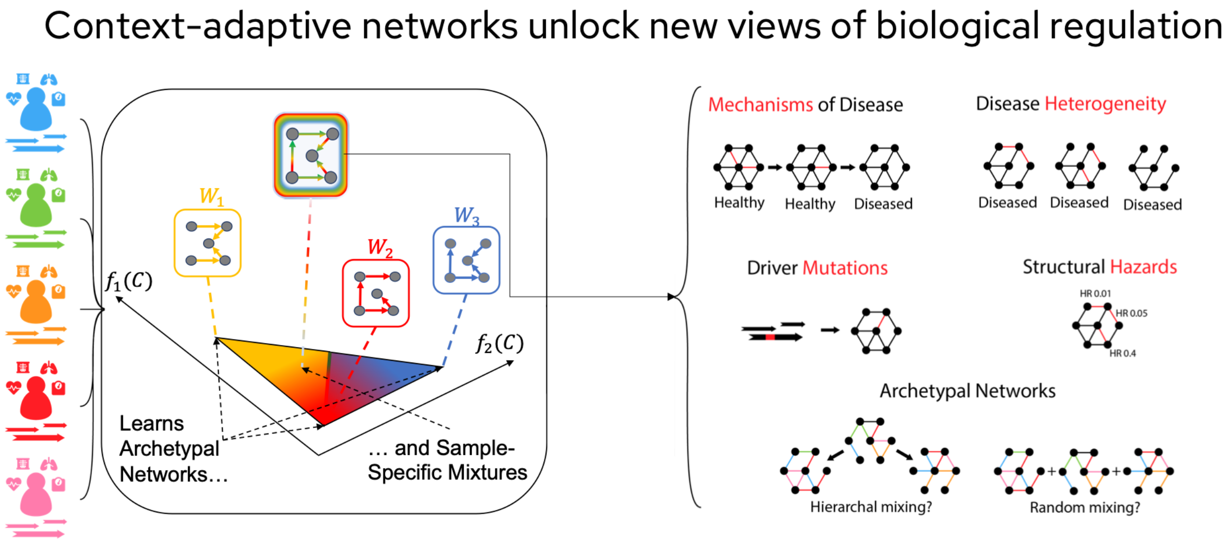

Context-adaptive networks unlock new views of biological regulation

Figure 11:

Performance Evaluation

Survey of Tools

Selection and Usage Guidance

Future Trends and Opportunities with Foundation Models

A New Paradigm for Context-Adaptive Inference

Universal Context Encoders

Dynamic Adaptation Mechanisms

Bridging with Statistical and Causal Reasoning

Next-Generation Methods for Contextualized Adaptive Inference

Modular Fine-Tuning and Compositional Adaptation

In-Context Learning and Mechanistic Insights

Reliability, Calibration, and Context-Sensitive Evaluation

Expanding Frameworks with Foundation Models

Foundation Models as Context

Feature Extraction and Interpretation:

Contextualized Representations for Downstream Modeling:

Post-hoc Interpretability:

Recent Innovations and Outlook

FLAN-MoE

LMPriors

Mixture of In-Context Experts

Open Problems

Open Research Questions

Can Reusable Modules Enable Portability Across Tasks?

What Are the Theoretical and Practical Benefits of Explicit Structure?

At What Level of Abstraction Should Explicit Structure Be Imposed?

What Theoretical and Practical Barriers Remain?

Interpretable-by-Design vs Post-hoc Interpretability: What Is the Right Path Forward?

Broader Challenges and Future Outlook

Conclusion

Overview of Insights

Context-Aware Efficiency: A Unifying Framework

Future Directions

Theoretical Foundations

Modular and Compositional Methods

Evaluation and Reliability

Responsible and Sustainable Deployment

References

Transformers learn in-context by gradient descent

What Can Transformers Learn In-Context? A Case Study of Simple Function Classes

Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers

Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Neural Tangents: Fast and Easy Infinite Neural Networks in Python

Statistical methods with varying coefficient models

A Survey of Deep Meta-Learning

LoRA: Low-Rank Adaptation of Large Language Models

Foundational Models Defining a New Era in Vision: A Survey and Outlook

A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT

Varying-Coefficient Models

Bayesian Edge Regression in Undirected Graphical Models to Characterize Interpatient Heterogeneity in Cancer

Statistical estimation in varying coefficient models

Time-Varying Coefficient Model Estimation Through Radial Basis Functions

Contextual Explanation Networks

Contextualized Machine Learning

Contextualized: Heterogeneous Modeling Toolbox

Learning to Estimate Sample-specific Transcriptional Networks for 7000 Tumors

NOTMAD: Estimating Bayesian Networks with Sample-Specific Structures and Parameters

Contextualized Policy Recovery: Modeling and Interpreting Medical Decisions with Adaptive Imitation Learning

Automated interpretable discovery of heterogeneous treatment effectiveness: A COVID-19 case study

Discriminative Subtyping of Lung Cancers from Histopathology Images via Contextual Deep Learning

Contextual Feature Selection with Conditional Stochastic Gates

Estimating time-varying networks

When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction

Learning Sample-Specific Models with Low-Rank Personalized Regression

Sketch-Based Anomaly Detection in Streaming Graphs

The Design of Experiments

Nouvelles méthodes pour la détermination des orbites des comètes

Theoria motus corporum coelestium in sectionibus conicis solem ambientium

Generalized Linear Models

Generalized Linear Models

Application of the Logistic Function to Bio-Assay

The Regression Analysis of Binary Sequences

Statistical Methods for Research Workers

Herd effects on the growth of beef bulls from different sources tested together under grazing conditions

Construct validity in psychological tests.

Estimation of Genetic Parameters

A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix

Recovery of inter-block information when block sizes are unequal

Random-Effects Models for Longitudinal Data

Estimation in generalized linear models with random effects

Approximate Inference in Generalized Linear Mixed Models

Bayes Estimates for the Linear Model

Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images

Bayesian Data Analysis

Functional Data Analysis

Generalized Additive Models

Representation Learning: A Review and New Perspectives

Multitask Learning

A Survey on Transfer Learning

Generalizing from a Few Examples

Matching Networks for One Shot Learning

Prototypical Networks for Few-shot Learning

A contextual-bandit approach to personalized news article recommendation

A New Approach to Linear Filtering and Prediction Problems

Online Learning: A Comprehensive Survey

Online Learning and Online Convex Optimization

A survey on concept drift adaptation

Learning with Drift Detection

Early Drift Detection Method

New ensemble methods for evolving data streams

Some aspects of the sequential design of experiments

Reinforcement Learning: An Introduction

Finite-time Analysis of the Multiarmed Bandit Problem

A Markovian Decision Process

Learning from Delayed Rewards

Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning

Neuronlike adaptive elements that can solve difficult learning control problems

Human-level control through deep reinforcement learning

Compression, restoration, resampling, ‘compressive sensing’: fast transforms in digital imaging

Speech Analysis and Synthesis by Linear Prediction of the Speech Wave

A vector space model for automatic indexing

Contextures: Representations from Contexts

Auto-Encoding Variational Bayes

Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Learning Transferable Visual Models From Natural Language Supervision

On the Opportunities and Risks of Foundation Models

A Survey on In-context Learning

Adaptive Mixtures of Local Experts

WILDS: A Benchmark of in-the-Wild Distribution Shifts

Continuous Temporal Domain Generalization

LFME: A Simple Framework for Learning from Multiple Experts in Domain Generalization

Scalable Multi-Domain Adaptation of Language Models using Modular Experts

Towards Modular LLMs by Building and Reusing a Library of LoRAs

Mixture of LoRA Experts

Optimal pointwise adaptive methods in nonparametric estimation

The Weighted Majority Algorithm

Selective Test-Time Adaptation for Unsupervised Anomaly Detection using Neural Implicit Representations

Test-Time Adaptation Induces Stronger Accuracy and Agreement-on-the-Line

Harder Tasks Need More Experts: Dynamic Routing in MoE Models

Unsupervised Learning via Meta-Learning

Bayesian scaling laws for in-context learning

An overview of statistical learning theory

A Closer Look into Mixture-of-Experts in Large Language Models

Shortcut Learning in Deep Neural Networks

Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization

The Selective Labels Problem

Model Selection and Estimation in Regression with Grouped Variables

Regression Shrinkage and Selection Via the Lasso

Data Analysis Using Regression and Multilevel/Hierarchical Models

Optimization Methods for Large-Scale Machine Learning

Contextual Bandits with Cross-learning

Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers

Adam: A Method for Stochastic Optimization

Language Models are Few-Shot Learners

Emergent Abilities of Large Language Models

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Scaling Laws for Neural Language Models

Training Compute-Optimal Large Language Models

Publication Trends on the Varying Coefficients Model: Estimating the Actual (Under)Utilization of a Highly Acclaimed Method for Studying Statistical Interactions

Covariance Selection

Graphical Models

High-dimensional graphs and variable selection with the Lasso

Sparse inverse covariance estimation with the graphical lasso

Joint estimation of multiple graphical models

The Joint Graphical Lasso for Inverse Covariance Estimation Across Multiple Classes

Fast spatio-temporally varying coefficient modeling with reluctant interaction selection

Spatially Varying Coefficient Models for Estimating Heterogeneous Mixture Effects

Network Varying Coefficient Model

Bayesian Inference for General Gaussian Graphical Models With Application to Multivariate Lattice Data

Bayesian Inference of Multiple Gaussian Graphical Models

Bayesian covariate-dependent graph learning with a dual group spike-and-slab prior

Tree Boosted Varying Coefficient Models

A tree-based varying coefficient model

VCBART: Bayesian Trees for Varying Coefficients

Penalized Spline Estimation for Varying-Coefficient Models

Deep Multimodal Learning with Missing Modality: A Survey

Domain Adaptation under Missingness Shift

Variational Autoencoder with Arbitrary Conditioning

GAIN: Missing Data Imputation using Generative Adversarial Nets

A class of pattern-mixture models for normal incomplete data

Multiple imputation of incomplete multilevel data using Heckman selection models

XGBoost: A Scalable Tree Boosting System

The Missing Indicator Method: From Low to High Dimensions

Recurrent Neural Networks for Multivariate Time Series with Missing Values

BRITS: Bidirectional Recurrent Imputation for Time Series

XGBoost

An Overview of Multi-Task Learning in Deep Neural Networks

Attention Is All You Need

Auto-Encoding Variational Bayes

Meta-Learning in Neural Networks: A Survey

MetaICL: Learning to Learn In Context

Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers

Transformers as Support Vector Machines

An Explanation of In-context Learning as Implicit Bayesian Inference

In-Context Learning Strategies Emerge Rationally

In-context Learning and Induction Heads

Learning without training: The implicit dynamics of in-context learning

What learning algorithm is in-context learning? Investigations with linear models

Transformers learn in-context by gradient descent

What Can Transformers Learn In-Context? A Case Study of Simple Function Classes

Can Transformers Learn Full Bayesian Inference in Context?

TabICL: A Tabular Foundation Model for In-Context Learning on Large Data

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Explainable AI: A Review of Machine Learning Interpretability Methods

Language Models as Knowledge Bases?

Towards A Rigorous Science of Interpretable Machine Learning

Rethinking Explainable Machine Learning as Applied Statistics

"Why Should I Trust You?": Explaining the Predictions of Any Classifier

A Unified Approach to Interpreting Model Predictions

Axiomatic Attribution for Deep Networks

Learning Important Features Through Propagating Activation Differences

This Looks Like That: Deep Learning for Interpretable Image Recognition

Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning

Understanding Black-box Predictions via Influence Functions

Inference Suboptimality in Variational Autoencoders

beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework

Deep Variational Information Bottleneck

Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)

Towards Automatic Concept-based Explanations

Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

A Framework for the Quantitative Evaluation of Disentangled Representations

Understanding intermediate layers using linear classifier probes

What do you learn from context? Probing for sentence structure in contextualized word representations

Locating and Editing Factual Associations in GPT

Knowledge Neurons in Pretrained Transformers

In-Context Explainers: Harnessing LLMs for Explaining Black Box Models

A Framework for Evaluating Post Hoc Feature-Additive Explainers

Evaluation of post-hoc interpretability methods in time-series classification

Concept Bottleneck Models

Manipulating and Measuring Model Interpretability

Principles of risk minimization for learning theory

Invariant Risk Minimization

The Risks of Invariant Risk Minimization

Out-of-Distribution Generalization via Risk Extrapolation (REx)

Conditional Variance Penalties and Domain Shift Robustness

Causal inference using invariant prediction: identification and confidence intervals

Towards Deep Learning Models Resistant to Adversarial Attacks

Robustness May Be at Odds with Accuracy

Adversarial Examples Are Not Bugs, They Are Features

The Rise and Potential of Large Language Model Based Agents: A Survey

Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization

Just Train Twice: Improving Group Robustness without Training Group Information

Environment Inference for Invariant Learning

Multilevel Statistical Models

Multilevel growth curve models that incorporate a random coefficient model for the level 1 variance function

A Bayesian multilevel time-varying framework for joint modeling of hospitalization and survival in patients on dialysis

Dynamic effects of increasing heterogeneity in financial markets

Bayesian Forecasting in Economics and Finance: A Modern Review

Bayesian Dynamic Factor Models for High-dimensional Matrix-valued Time Series

International Asset Allocation With Regime Shifts

A Model-Based Method for Remaining Useful Life Prediction of Machinery

Predictive maintenance in the Industry 4.0: A systematic literature review

Predictive Maintenance Approaches in Industry 4.0: A Systematic Literature Review

A data-driven and context-aware approach for demand forecasting in the beverage industry

Chaotic Bayesian Inference: Strange Attractors as Risk Models for Black Swan Events

A Multi-target Bayesian Transformer Framework for Predicting Cardiovascular Disease Biomarkers during Pandemics

Bayesian Dynamic Factor Models for High-dimensional Matrix-valued Time Series

Bayesian Models for Joint Selection of Features and Auto-Regressive Lags: Theory and Applications in Environmental and Financial Forecasting

SafeInfer: Context Adaptive Decoding Time Safety Alignment for Large Language Models

Robustness, Evaluation and Adaptation of Machine Learning Models in the Wild

A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions

Retrieval-Augmented Generation for AI-Generated Content: A Survey

Billion-scale similarity search with GPUs

From Human Memory to AI Memory: A Survey on Memory Mechanisms in the Era of LLMs

Memory OS of AI Agent

Efficient Streaming Language Models with Attention Sinks

Efficient Memory Management for Large Language Model Serving with PagedAttention

On the Dangers of Stochastic Parrots

Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer

Curvature-Torsion Entropy for Twisted Curves under Curve Shortening Flow

LMPriors: Pre-Trained Language Models as Task-Specific Priors

AdapterFusion: Non-Destructive Task Composition for Transfer Learning

Does Combining Parameter-efficient Modules Improve Few-shot Transfer Accuracy?

In-Context Learning through the Bayesian Prism

On Calibration of Modern Neural Networks

Holistic Evaluation of Language Models

GPT-4 Technical Report

The Llama 3 Herd of Models

TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

CHiLL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models

Learning Interpretable Style Embeddings via Prompting LLMs

Tree Prompting: Efficient Task Adaptation without Fine-Tuning

One Embedder, Any Task: Instruction-Finetuned Text Embeddings

Augmenting interpretable models with large language models during training

Explaining Datasets in Words: Statistical Models with Natural Language Parameters

Mixture-of-Experts Meets Instruction Tuning:A Winning Combination for Large Language Models

Investigating Lane-Free Traffic with a Dynamic Driving Simulator

An Overview of Perception Methods for Horticultural Robots: From Pollination to Harvest

Agentic Context Engineering: Evolving Contexts for Self-Improving Language Models

Appendix A

A.0 Preliminaries and identities

- **Joint features.** (x, c)

$$\psi(x, c) := x \otimes \phi(c) \in \mathbb{R}^{d_x d_c}.$$

a (i, j) $\psi_a := \psi(x_a, c_a)$

- **Design/labels/weights.** $N = \sum_i m_i$

$$Z \in \mathbb{R}^{N \times d_x d_c} \text{ with rows } Z_a = \psi_a^T, \qquad y \in \mathbb{R}^N, \qquad W = \mathrm{diag}(w_a) \in \mathbb{R}^{N \times N},\ w_a \geq 0.$$

$$K := ZZ^\top$$

$$K_W := W^{1/2} K\, W^{1/2} \ = \ W^{1/2} Z Z^\top W^{1/2}.$$

(x, c) $k(\cdot, (x, c)) := Z\,\psi(x, c) \in \mathbb{R}^N$ $k_{(x,c)} := W^{1/2} k(\cdot, (x, c))$

- **Vectorization identity.** A, B, C

$$\mathrm{vec}(ABC) = \big(C^\top \otimes A\big)\mathrm{vec}(B), \quad \langle \mathrm{vec}(B),\, x \otimes z\rangle = x^\top B z.$$

- **Weighted ridge solution.** $X \in \mathbb{R}^{N \times p}$

$$\min_\beta \ \|W^{1/2}(y - X\beta)\|_2^2 + \lambda \|\beta\|_2^2$$

$$\widehat{\beta} = (X^\top W X + \lambda I)^{-1} X^\top W y$$

$$\widehat{\beta} = X^\top W^{1/2} \big(W^{1/2} X X^\top W^{1/2} + \lambda I\big)^{-1} W^{1/2} y.$$

$x_\star$

$$\widehat{f}\,(x_\star) = x_\star^\top \widehat{\beta} \ = \ \underbrace{\big(W^{1/2} X x_\star\big)^\top}_{k_\star^\top} \big(W^{1/2} X X^\top W^{1/2} + \lambda I\big)^{-1} W^{1/2} y.$$

kernel ridge regression $K_W = W^{1/2} X X^\top W^{1/2}$

$k_\star = W^{1/2} X x_\star$

A.1 Proof of Proposition 1(A): explicit varying-coefficients ⇔ weighted KRR on joint features

$y = \langle \theta(c), x\rangle + \varepsilon$ $\mathbb{E}[\varepsilon] = 0$

$\theta(c) = B\,\phi(c)$ $B \in \mathbb{R}^{d_x \times d_c}$ $\lambda \|B\|_F^2$

Step 1 (reduce to ridge in joint-feature space).

B $\beta = \mathrm{vec}(B) \in \mathbb{R}^{d_x d_c}$

$$x_a^T B\,\phi(c_a) = \langle \beta,\, x_a \otimes \phi(c_a)\rangle = \langle \beta,\, \psi_a\rangle.$$

$$\min_{\beta\in\mathbb{R}^{d_x d_c}} \;\|W^{1/2}\big(y - Z\beta\big)\|_2^2 + \lambda\|\beta\|_2^2,$$

$$X \equiv Z$$

Step 2 (closed form and prediction).

$$\widehat{\beta} = (Z^T W Z + \lambda I)^{-1} Z^T W y,$$

(x, c) $\psi(x, c)$

$$\hat{y}(x,c) = \psi(x,c)^T\widehat{\beta} = \underbrace{\big(W^{1/2} Z\,\psi(x,c)\big)}_{k_{(x,c)}}^{\top}\big(W^{1/2}ZZ^{\top}W^{1/2} + \lambda I\big)^{-1} W^{1/2} y.$$

Step 3 (kernel form).

$K := ZZ^T$ $K_W := W^{1/2} K W^{1/2}$

$$\boxed{\hat{y}(x,c) \;=\; k_{(x,c)}^T\big(K_W + \lambda I\big)^{-1} W^{1/2} y}.$$

(a, b) K

$$K_{ab} = \langle\psi_a, \psi_b\rangle = \big\langle x_a \otimes \phi(c_a),\, x_b \otimes \phi(c_b)\big\rangle = \langle x_a, x_b\rangle \cdot \langle\phi(c_a), \phi(c_b)\rangle,$$

KRR on joint features W

A.2 Proof of Proposition 1(B): linear ICL ⇒ kernel regression

$S(c)$ **linear**

$$q(x,c) = Q\,\psi(x,c), \qquad k_a = K\,\psi_a, \qquad v_a = V\,\psi_a,$$

$Q \in \mathbb{R}^{d_q\times d_\psi}$ $K \in \mathbb{R}^{d_k\times d_\psi}$ $V \in \mathbb{R}^{d_v\times d_\psi}$ $d_\psi = d_x d_c$ **unnormalized**

a

$$s_a(x,c) := w_a\,\langle q(x,c), k_a\rangle \;=\; w_a\,\psi(x,c)^T Q^T K\,\psi_a.$$

$\alpha_a(x,c) := s_a(x,c)/\sum_b s_b(x,c)$

$\{\alpha_a\}$

$$z(x,c) = \sum_a \alpha_a(x,c)\,v_a, \qquad \hat{y}(x,c) = u^T z(x,c).$$

(B1)

(B2)

A.2.1 (B1) Fixed attention, trained linear head ⇒ exact KRR

Q, K, V $\alpha_a(x,c)$ **deterministic**

(x, c) **feature map**

$$\varphi(x,c) := \sum_a \alpha_a(x,c)\,v_a \;\in\; \mathbb{R}^{d_v}.$$

$\varphi_a := \varphi(x_a, c_a)$ $\Phi \in \mathbb{R}^{N\times d_v}$ u

$$\widehat{u} \in \arg\min_u \;\|W^{1/2}(y - \Phi u)\|_2^2 + \lambda\|u\|_2^2$$

$\widehat{u} = (\Phi^T W\Phi + \lambda I)^{-1}\Phi^T W y$

$$\hat{y}(x,c) = \varphi(x,c)^T\hat{u} = \underbrace{\big(W^{1/2}\Phi\,\varphi(x,c)\big)}_{k_{(x,c)}}^{T}\big(W^{1/2}\Phi\Phi^T W^{1/2} + \lambda I\big)^{-1}W^{1/2}y.$$

$$\boxed{\hat{y}(x,c) = k_{(x,c)}^T\big(K_W + \lambda I\big)^{-1}W^{1/2}y}, \quad K_W := W^{1/2}\underbrace{(\Phi\Phi^T)}_{=:K}W^{1/2},$$

kernel ridge regression

$$k\big((x,c),(x',c')\big) = \langle\varphi(x,c),\varphi(x',c')\rangle.$$

$v_a = V\psi_a$ $\alpha_a(x,c) \propto w_a\,\psi(x,c)^T Q^T K\psi_a$ φ **weighted average of joint features** $\{\psi_a\}$

A.2.2 (B2) Training attention in the linearized/NTK regime ⇒ kernel regression with NTK

$\theta = (Q,K,V,u)$

θ_0 **linearized model** θ_0

$$\hat{y}_\theta(x,c) \;\approx\; \hat{y}_{\theta_0}(x,c) + \nabla_\theta\hat{y}_{\theta_0}(x,c)^T(\theta-\theta_0) =: \hat{y}_{\theta_0}(x,c) + \phi_{\text{NTK}}(x,c)^T(\theta-\theta_0),$$

$\phi_{\text{NTK}}(x,c) := \nabla_\theta\hat{y}_{\theta_0}(x,c)$ **tangent features**

kernel regression with the NTK

$$k_{\text{NTK}}\big((x,c),(x',c')\big) := \big\langle\phi_{\text{NTK}}(x,c),\,\phi_{\text{NTK}}(x',c')\big\rangle,$$

K_{NTK}

ϕ_{NTK} **linear attention**

linear transforms of joint features $\hat{y}(x,c) = u^T\sum_a \alpha_a(x,c)\,V\psi_a$ θ_0

- **Readout path (u).** $\partial\hat{y}/\partial u = \sum_a \alpha_a(x,c)\,V\psi_a = \varphi_0(x,c)$ $\{\psi_a\}$
- **Value path (V).** $\partial\hat{y}/\partial V = \sum_a \alpha_a(x,c)\,u\,\psi_a^T$ $(u\otimes I)\sum_a \alpha_a(x,c)\psi_a$ $\{\psi_a\}$
- **Query/key paths (Q,K).** $s_a = w_a\,\psi(x,c)^T Q^T K\psi_a$ $\alpha_a = s_a/\sum_b s_b$ α_a Q K $\psi(x,c)$ $\{\psi_a\}$

$$\frac{\partial\alpha_a}{\partial Q} \propto \sum_b \big[\delta_{ab} - \alpha_b(x,c)\big]\,w_a w_b\big(K\psi_a\,\psi(x,c)^T\big),$$

$$\frac{\partial\alpha_a}{\partial K} \propto \sum_b \big[\delta_{ab} - \alpha_b(x,c)\big]\,w_a w_b\big(\psi(x,c)\,\psi_a^T Q^T\big),$$

$\partial\hat{y}/\partial Q$ $\partial\hat{y}/\partial K$ $\psi(x,c)$ ψ_a u V $\psi(x,c)$ $\{\psi_a\}$

$$\phi_{\mathrm{NTK}}(x,c) = \mathcal{L}\big(\psi(x,c), \{\psi_a\}\big),$$

$\mathcal{L}$ θ_0 W

dot-product

$$k_{\mathrm{NTK}}\big((x,c),(x',c')\big) = \Psi(x,c)^T\,\mathcal{M}\,\Psi(x',c'),$$

$\mathcal{M}$ Ψ

$\psi(x,c)$ $\{\psi_a\}$ k_{NTK}

linear transforms of the joint features

Assumptions for A.2.2.

A.3 Proof of Corollary 1: retrieval/gating/weighting as kernel/measure choices

$$\hat{y}(x,c) = k^T_{(x,c)}\big(K^\sharp + \lambda I\big)^{-1}\mu,$$

$K^\sharp$

$K_W = W^{1/2}ZZ^TW^{1/2}$ $W^{1/2}\Phi\Phi^TW^{1/2}$ K_{NTK} $k_{(x,c)}$

$\mu = W^{1/2}y$

- **Retrieval $R(c)$ / gating.** $S(c)$ **removes or adds rows/columns** $K^\sharp$ $k_{(x,c)}$ **empirical measure**
- **Weights $w_{ij}(c)$.** W K $K_W = W^{1/2}KW^{1/2}$ k $k_{(x,c)} = W^{1/2}k$ **importance weighting**
- **Induced kernels.** **feature map** $\psi \mapsto V\psi$ $\psi \mapsto Q\psi$ $k((x,c),(x',c')) = \langle \Phi(x,c), \Phi(x',c')\rangle$

neighborhood selection

kernel choice

A.4 Remarks

- **No Gaussianity is required.** $y = f(x,c) + \varepsilon$ $\mathbb{E}[\varepsilon] = 0$
- **Early stopping vs. explicit ridge.** λ
- **Multiple layers / nonlinear value stacks.**