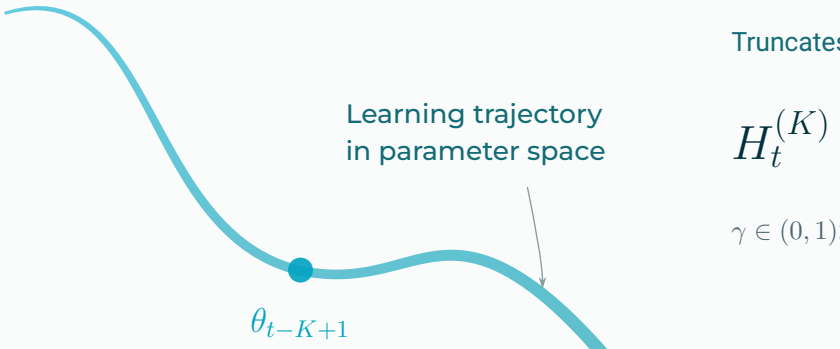
Matrix-Free Natural Gradient via Stochastic Truncated Metrics

Accumulate K stochastic rank-one estimators; apply the metric inverse through a small $K \times K$ solve.

Setting A classifier network $p_\theta(y \mid x)$ with parameters $\theta \in \mathbb{R}^d$, $x \in \mathcal{D}_t$ (a batch of B samples), $y \in \{1, \dots, C\}$.

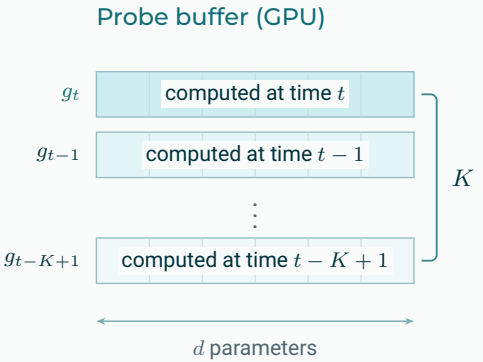


Stochastic Truncated Metric (STM)

Truncates a damped stochastic FIM EMA to the latest $K \ll d$ steps.

$$H_t^{(K)} = \gamma I + (1-\gamma)(1-\beta) \sum_{j=0}^{K-1} \beta^j g_{t-j} g_{t-j}^\mathsf{T}$$

$\gamma \in (0,1)$: floor; $\beta \in [0,1)$: decay; K: probe count.



Applying the STM inverse is matrix-free

(no dense $d \times d$ or layer-wise matrix is constructed or inverted).

$$[H_t^{(K)}]^{-1} v = \frac{v - U(\gamma I_K + U^\mathsf{T} U)^{-1} U^\mathsf{T} v}{\gamma}$$

Woodbury: $K \times K$ solve — v: training gradient

$$U = [\sqrt{(1-\gamma)(1-\beta)\beta^j}\, g_{t-j}]_{j=0}^{K-1} \in \mathbb{R}^{d \times K}$$

Auxiliary **space** $O(Kd + K^2)$

Update **time** $O(Kd + K^3)$

Stochastic rank-one unbiased estimation of the Fisher information matrix (FIM)

ε_{xy}: independent Rademacher

$$h_t(\theta) = \frac{2}{\sqrt{B}} \sum_{x \in \mathcal{D}_t} \sum_y \varepsilon_{xy} \sqrt{p_\theta(y \mid x)}$$

$$g_t = \nabla_\theta h_t(\theta_t) = \frac{1}{\sqrt{B}} \sum_{x \in \mathcal{D}_t} \sum_y \varepsilon_{xy} \sqrt{p_{\theta_t}(y \mid x)}\, \nabla_\theta \log p_{\theta_t}(y \mid x)$$

requires one additional backward pass

$$\mathbb{E}_\varepsilon[g_t g_t^\mathsf{T}] = \frac{1}{B} \sum_{x \in \mathcal{D}_t} \sum_y p_{\theta_t}(y \mid x) \nabla_\theta \log p_{\theta_t}(y \mid x)\, \nabla_\theta \log p_{\theta_t}(y \mid x)^\mathsf{T} = \bar{\mathcal{F}}_B(\theta_t)$$

unbiased FIM estimator

Paper

On two benchmarks, STM's mean test accuracy was within **0.7 percentage points of K-FAC**, with comparable or lower per-epoch time.