Limitations of the empirical Fisher approximation for natural gradient descent

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Max Planck Institute for Intelligent Systems

Frederik Kunstner, Lukas Balles, Philipp Hennig

Abstract

Natural gradient descent uses the Fisher information matrix to adapt to the geometry. Several works advocate for the empirical Fisher approximation and draw connections between second-order methods and heuristics like Adam. We show that this approximation does not capture the problem geometry.

Natural Gradient Descent

Goal: learn the conditional distribution $y \mid x$;

$$\mathcal{L}(\theta) = -\sum_n \log p_\theta(y_n|x_n)$$

Natural Gradient Descent: preconditioned gradient update with the Fisher information F,

$$\theta_{t+1} = \theta_t - F(\theta_t)^{-1}\nabla\mathcal{L}(\theta_t)$$

The landscape of Fisher matrices

The Fisher of the joint $p_\theta(x,y) = p(x)p_\theta(y|x)$ is

$$N\mathbb{E}_{p(x)}\mathbb{E}_{p_\theta(y|x)}\left[\nabla \log p_\theta(y|x)\nabla \log p_\theta(y|x)^\top\right]$$

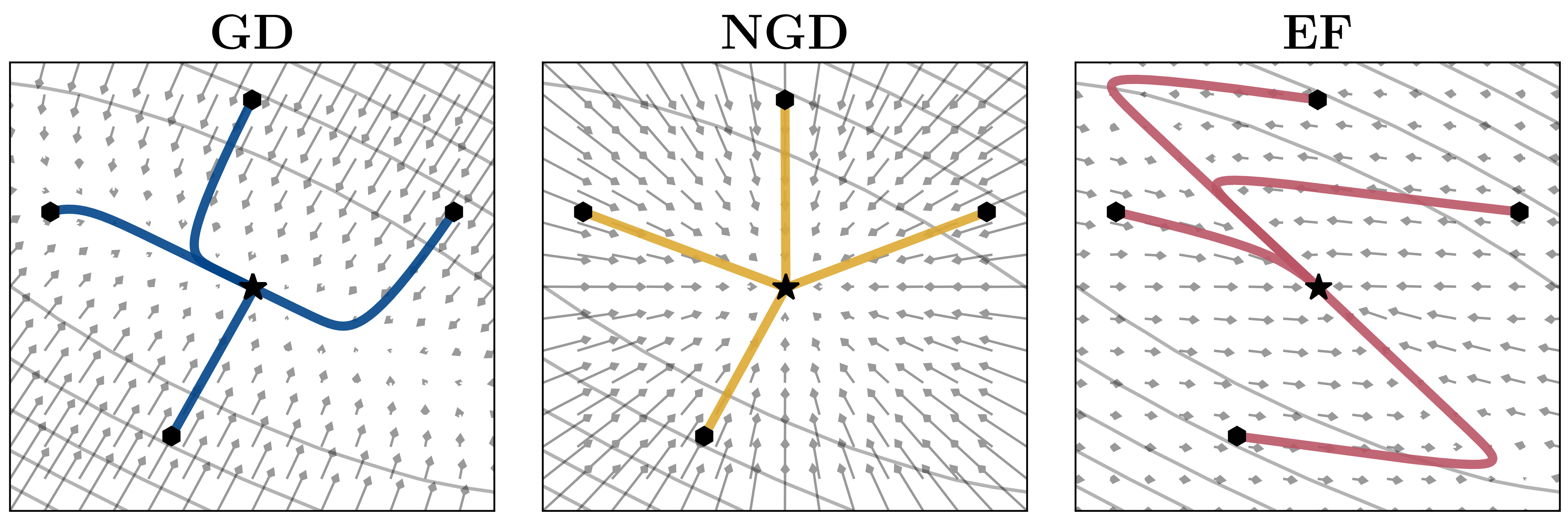
If $p(x)$ is unknown, the Fisher of the conditional $p_\theta(y|x_n)$ (empirical x_n) also works;

$$\sum_n \mathbb{E}_{p_\theta(y|x_n)}\left[\nabla \log p_\theta(y|x_n)\nabla \log p_\theta(y|x_n)^\top\right]$$

But the empirical Fisher approximaton uses the empirical y_n;

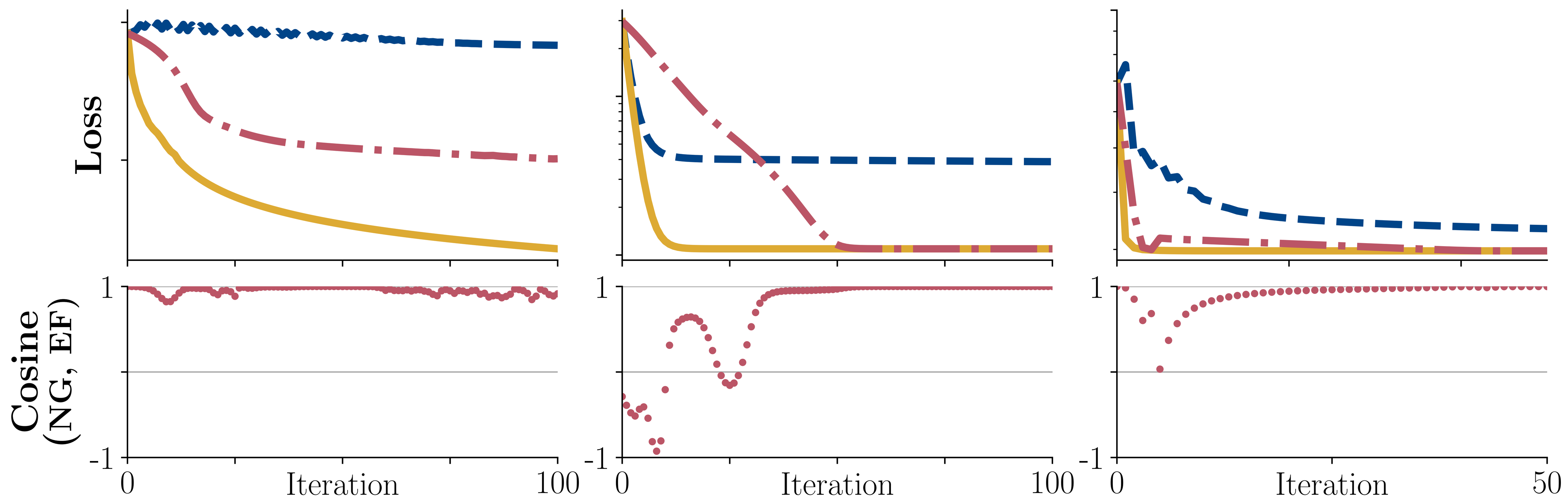
$$\sum_n \nabla \log p_\theta(y_n|x_n)\nabla \log p_\theta(y_n|x_n)^\top$$

The empirical Fisher is a bad preconditioner ...



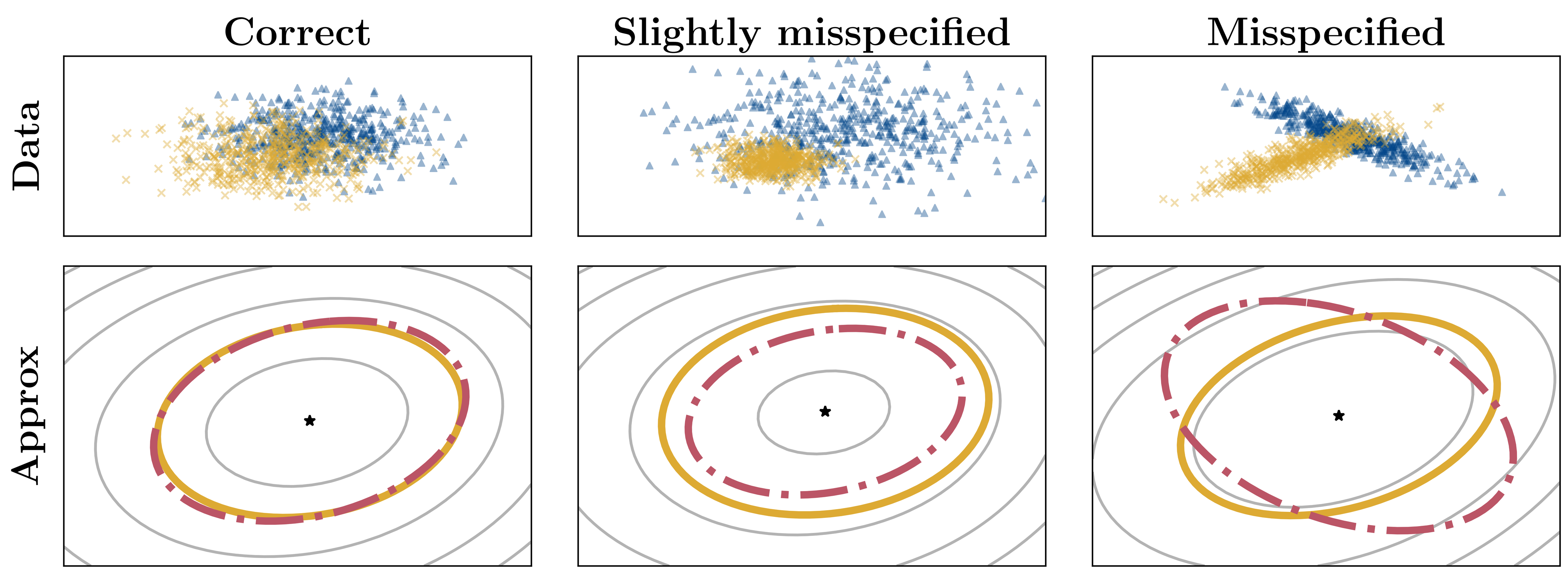
- The problem is ill conditioned; Gradient descent struggles and natural gradients adapt to the geometry
- EF distorts direction and magnitude; large gradient $\implies$ small update small gradient $\implies$ large update

... which leads to bad behavior during optimization ...



- Small update when gradient is large $\implies$ step-size tuning is hard
- Even if tuned, direction might fail; can be opposite of natural gradient

... and can misestimate curvature, even at the minimum



- EF should $\approx$ Fisher at the minimum if model is well-specified and there is enough data $\implies$ hard to check in advance.
- Large models can help; more likely to be well-specified but also need more data

kunstner@cs.ubc.ca, lballes@tue.mpg.de, ph@tue.mpg.de