Question 1. *Provide a summary of the contributions of this paper.*

Response:

MAP –

posterior dist. after samples D

loss function when training only B

– max likelihood estimation

Question 2. *Derive equation (2). Your response should point out any assumptions the derivation is making.*

doesnt depend on θ

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}_B|\theta) + \log p(\theta|\mathcal{D}_A) - \log p(\mathcal{D}_B) \tag{2}$$

params dataset

posterior on θ given D_A prior on θ for training on D_B

Response:

Bayes: $$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$

$$\log P(\theta|D) = \log P(D|\theta) + \log P(\theta) - \log P(D)$$

Assume $D = D_A \cup D_B$, independent samples

So $$\log P(D|\theta) = \log P(D_A|\theta) + \log P(D_B|\theta)$$

$$\log P(\theta|D) = \log P(D_A|\theta) + \log P(D_B|\theta) + \log P(\theta) - \log P(D_A) - \log P(D_B)$$

$$= \log P(D_B|\theta) + \log P(\theta|D_A) - \log P(D_B)$$

θ_A^* – params. that were result of training on A

Question 3. *Explain how formulation (3) is obtained from equation (2).*

parameters

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2 \tag{3}$$

$P(\theta | D_A)$ is intractible – use Laplace approx

① Approximate $\log P(\theta | D_A)$ as $\underbrace{P(\theta_A^* | D_A) - (\theta - \theta_A^*) H (\theta - \theta_A^*)}_{\text{Taylor Expansion}}$

Hession $\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}$

Note: H approximated by Fisher Infor. Matrix

Only use diagonal elements of FIM (ignoring covariance terms)

Modeled posterior locally as Gaussian which has all components as independent

Why only consider the diagonal? Computing FIM is intractable

Defn: Fisher information matrix of P_θ

$$F = \mathbb{E}_{z \sim P_\theta} D_\theta^2 \log P_\theta(z) = - \mathbb{E}_{z \sim P_\theta} \left[\nabla \log P_\theta(z) \, \nabla \log P_\theta(z)^t \right]$$

$$f(\theta) \approx f(\theta_*) + \nabla_\theta f(\theta_*)\cdot(\theta-\theta_*) + \frac{1}{2}(\theta-\theta_*)^t H (\theta-\theta_*) + \dots$$

$\mathbb{R}^1$ $\quad \in \mathbb{R}^d$

$$H_{ij} = \frac{\partial^2 f}{\partial\theta_i \partial\theta_j}$$

Consider a probability dist. that depends on θ

$$p(z|\theta), \quad P_\theta(z)$$

How much does θ affect P_θ on average? Consider $\theta \in \mathbb{R}^d$

$$\mathop{\mathbb{E}}_{z\sim P_\theta} \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log P_\theta(z)$$

Evaluate 2^{nd} derivative:

$$\frac{\partial^2}{\partial\theta_j \partial\theta_i} \log P_\theta(z) = \frac{\partial}{\partial\theta_j} \frac{\partial_{\theta_i} P_\theta(z)}{P_\theta(z)}$$

$$= \frac{\partial_{\theta_j}\partial_{\theta_i} P_\theta(z)}{P_\theta(z)} - \frac{\partial_{\theta_i} P_\theta(z)}{P_\theta(z)} \frac{\partial_{\theta_j} P_\theta(z)}{P_\theta(z)}$$

$$= \frac{\partial^2_{\theta_i\theta_j} P_\theta(z)}{P_\theta(z)} - \partial_{\theta_i} \log P_\theta(z)\, \partial_{\theta_j} \log P_\theta(z)$$

$$\mathop{\mathbb{E}}_{z\sim P_\theta} \partial^2_{\theta_i\theta_j} \log P_\theta(z) = \cancelto{0}{\mathop{\mathbb{E}}_{z\sim P_\theta} \frac{\partial^2_{\theta_i\theta_j} P_\theta(z)}{P_\theta(z)}} - \mathop{\mathbb{E}}_{z\sim P_\theta} \partial_{\theta_i} \log P_\theta(z)\, \partial_{\theta_j} \log P_\theta(z)$$

$$\mathop{\mathbb{E}}_{z\sim P_\theta} D^2 \log P_\theta(z) = -\mathop{\mathbb{E}}_{z\sim P_\theta} \left[\nabla \log P_\theta(z)\, \nabla \log P_\theta(z)^t \right]$$

Calc:

$$\mathop{\mathbb{E}}_{Z\sim P_\theta} \frac{\partial^2_{\theta_i\theta_j} P_\theta(Z)}{P_\theta(Z)} = \int_R \frac{\partial^2_{\theta_i\theta_j} P_\theta(Z)}{\cancel{P_\theta(Z)}} \cancel{P_\theta(Z)}\, dZ$$

$$= \partial^2_{\theta_i\theta_j} \int_R P_\theta(Z)\, dZ$$

$$= 0.$$

Similarly:

$$\mathop{\mathbb{E}}_{Z\sim P_\theta} \partial_{\theta_i} \log P_\theta(Z) = \mathop{\mathbb{E}}_{Z} \frac{\partial_{\theta_i} P_\theta(Z)}{P_\theta(Z)} = \int \frac{\partial_{\theta_i} P_\theta(Z)}{\cancel{P_\theta(Z)}} \cancel{P_\theta(Z)}\, dZ$$

$$= \partial_{\theta_i} \int P_\theta(Z)\, dZ =$$

$$= \partial_{\theta_i} 1 = 0.$$

Limitations of the Empirical Fisher Approximation for Natural Gradient Descent

Frederik Kunstner[1,2,3]
kunstner@cs.ubc.ca

Lukas Balles[2,3]
lballes@tue.mpg.de

Philipp Hennig[2,3]
ph@tue.mpg.de

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland[1]
University of Tübingen, Germany[2]
Max Planck Institute for Intelligent Systems, Tübingen, Germany[3]

Abstract

Natural gradient descent, which preconditions a gradient descent update with the Fisher information matrix of the underlying statistical model, is a way to capture partial second-order information. Several highly visible works have advocated an approximation known as the empirical Fisher, drawing connections between approximate second-order methods and heuristics like Adam. We dispute this argument by showing that the empirical Fisher—unlike the Fisher—does not generally capture second-order information. We further argue that the conditions under which the empirical Fisher approaches the Fisher (and the Hessian) are unlikely to be met in practice, and that, even on simple optimization problems, the pathologies of the empirical Fisher can have undesirable effects.

1 Introduction

Consider a supervised machine learning problem of predicting outputs $y \in \mathbb{Y}$ from inputs $x \in \mathbb{X}$. We assume a probabilistic model for the conditional distribution of the form $p_\theta(y|x) = p(y|f(x,\theta))$, where $p(y|\cdot)$ is an exponential family with natural parameters in $\mathbb{F}$ and $f \colon \mathbb{X} \times \mathbb{R}^D \to \mathbb{F}$ is a prediction function parameterized by $\theta \in \mathbb{R}^D$. Given N iid training samples $(x_n, y_n)_{n=1}^N$, we want to minimize

$$\mathcal{L}(\theta) := -\sum_n \log p_\theta(y_n|x_n) = -\sum_n \log p(y_n|f(x_n,\theta)). \tag{1}$$

This framework covers common scenarios such as least-squares regression ($\mathbb{Y} = \mathbb{F} = \mathbb{R}$ and $p(y|f) = \mathcal{N}(y; f, \sigma^2)$ with fixed σ^2) or C-class classification with cross-entropy loss ($\mathbb{Y} = \{1, \dots, C\}$, $\mathbb{F} = \mathbb{R}^C$ and $p(y = c|f) = \exp(f_c)/\sum_i \exp(f_i)$) with an arbitrary prediction function f. Eq. (1) can be minimized by gradient descent, which updates $\theta_{t+1} = \theta_t - \gamma_t \nabla \mathcal{L}(\theta_t)$ with step size $\gamma_t \in \mathbb{R}$. This update can be *preconditioned* with a matrix B_t that incorporates additional information, such as local curvature, $\theta_{t+1} = \theta_t - \gamma_t {B_t}^{-1} \nabla \mathcal{L}(\theta_t)$. Choosing B_t to be the Hessian yields Newton's method, but its computation is often burdensome and might not be desirable for non-convex problems. A prominent variant in machine learning is *natural gradient descent* [NGD; Amari, 1998]. It adapts to the *information geometry* of the problem by measuring the distance between parameters with the Kullback–Leibler divergence between the resulting distributions rather than their Euclidean distance, using the Fisher information matrix (or simply "Fisher") of the model as a preconditioner,

$$\mathrm{F}(\theta) := \sum_n \mathbb{E}_{p_\theta(y|x_n)} \left[\nabla_\theta \log p_\theta(y|x_n) \, \nabla_\theta \log p_\theta(y|x_n)^\top \right]. \tag{2}$$

While this motivation is conceptually distinct from approximating the Hessian, the Fisher coincides with a generalized Gauss-Newton [Schraudolph, 2002] approximation of the Hessian for the problems presented here. This gives NGD theoretical grounding as an approximate second-order method.

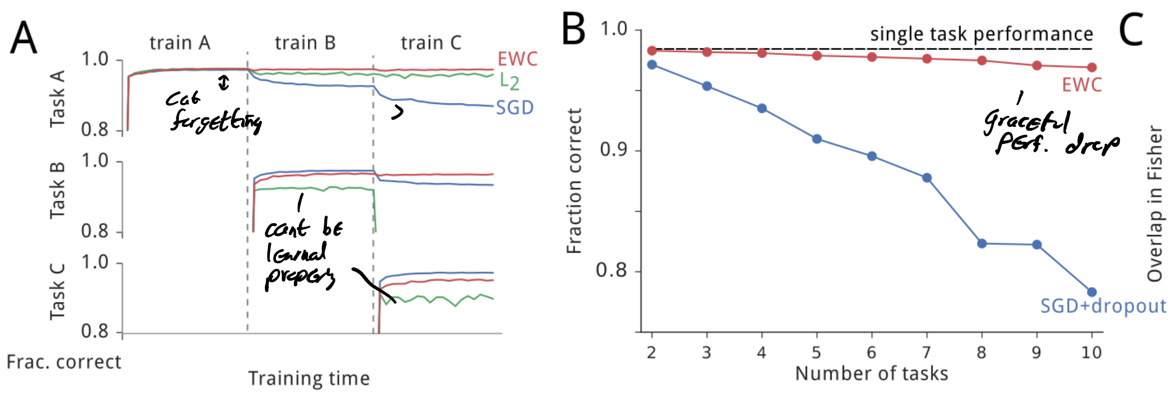
A number of recent works in machine learning have relied on a certain approximation of the Fisher, which is often called the *empirical Fisher (EF)* and is defined as

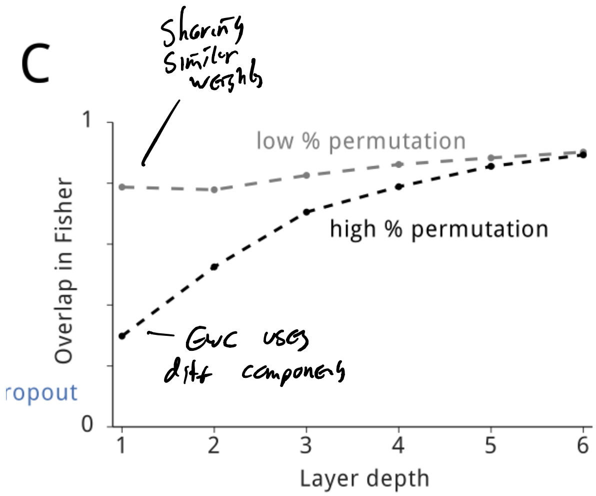
$$\widetilde{\mathrm{F}}(\theta) := \sum_n \nabla_\theta \log p_\theta(y_n|x_n) \, \nabla_\theta \log p_\theta(y_n|x_n)^\top. \tag{3}$$

Code available at github.com/fkunstner/limitations-empirical-fisher.

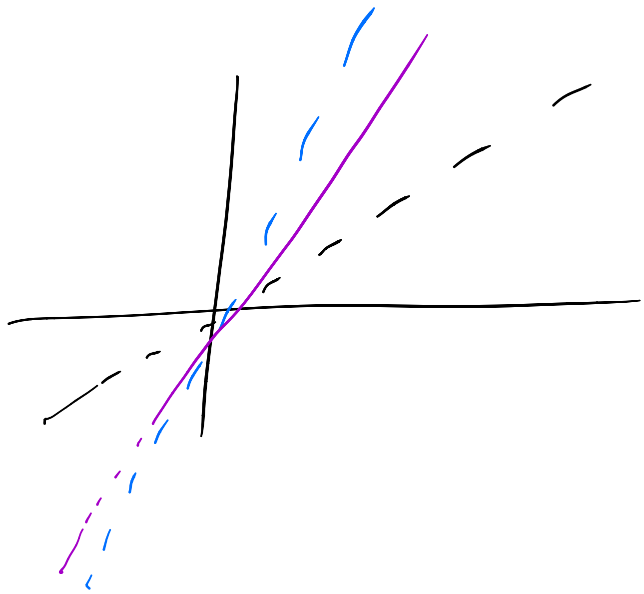
Question 4. *Explain Figure 2ab. Make sure to include the context, a statement of what literally is plotted, what is to be observed, and what is concluded.*

Question 5. *Explain Figure 2c. Make sure to include the context, a statement of what literally is plotted, what is to be observed, and what is concluded.*





Question 6. *How is the algorithm in this paper biologically inspired? Why is the method called 'elastic weight consolidation'?*



CS 7150: Deep Learning — Spring 2021 — Paul Hand

Day 19 — Preparation Questions For Class
Due: Wednesday 3/31/2021 at 2:30pm via Gradescope

Names: [Put The Names Of Your Group Here]

You may consult any and all resources in answering the questions. Your goal is to have answers that are ready to be shared with the class (or on a hypothetical job interview) as written. Your answers should be as concise as possible. When asked to explain a figure, your response should have the following structure: provide context (state what experiment was being run / state what problem is being solved), state what has been plotted, remark on what we observe from the plots, and interpret the results.

Submit one document for your group and tag all group members. We recommend you use Overleaf for joint editing of this TeX document.

Directions: Read 'Overcoming catastrophic forgetting in neural networks' by Kirkpatrick et al.

- Read Sections 1, 2.0, 2.1, 3

Question 1. *Provide a summary of the contributions of this paper.*

Response:

Question 2. *Derive equation (2). Your response should point out any assumptions the derivation is making.*

Response:

Question 3. *Explain how formulation (3) is obtained from equation (2).*

Response:

Question 4. *Explain Figure 2ab. Make sure to include the context, a statement of what literally is plotted, what is to be observed, and what is concluded.*

Response:
Context:
What is plotted:
What we observe:
Interpretation:

Question 5. *Explain Figure 2c. Make sure to include the context, a statement of what literally is plotted, what is to be observed, and what is concluded.*

Response:

Context:
What is plotted:
What we observe:
Interpretation:

Question 6. *How is the algorithm in this paper biologically inspired? Why is the method called 'elastic weight consolidation'?*

Response: