Fisher Information

Applications in Gradient Descent and Incremental Learning

Megh Shukla

RD I/CEI
Mercedes-Benz Research and Development India

January 6, 2022

Overview

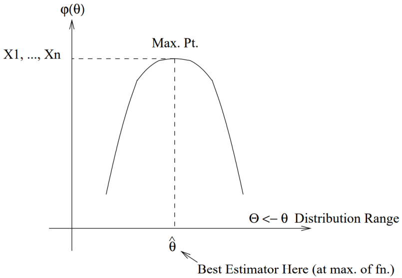
1. Why Fisher Information?
2. Applications - Natural Gradient
3. Applications - Online Learning
4. Applications - Active Learning

Revisiting Maximum Likelihood

Given some observations x, we want to obtain θ that maximizes $f(x|\theta)$.
With the *i.i.d* assumption, our likelihood function is $\psi(\theta) = f(x_1|\theta) \times \ldots \times f(x_n|\theta)$.

Maximum Likelihood Estimate

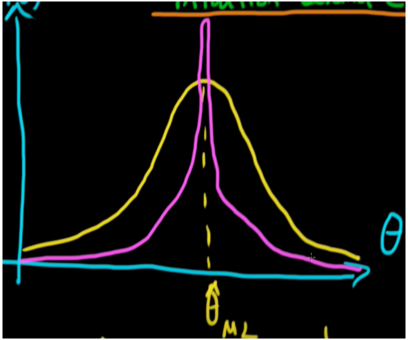
$$\psi(\hat{\theta}) = \arg\max_{\theta} \ \psi(\theta)$$



Log likelihood makes it easy to obtain $\hat{\theta} = \arg\max_{\theta} \ \psi(\theta) = \sum_{i=1}^{N} \log f(x_i|\theta)$

Revisiting Maximum Likelihood

Q1. So what after $\hat{\theta}$? How confident are we about our prediction?



Q2. Are we sure about $\hat{\theta} \to \theta_0$ as $n \to \infty$?

... Fisher Information to the rescue!

Definition

Fisher Information

$$\mathcal{I}_\theta = \mathbb{E}_x \left[\nabla_\theta \log p(x|\theta) \, \nabla_\theta \log p(x|\theta)^T \right]$$

. . .

. . .

. . .

. . .

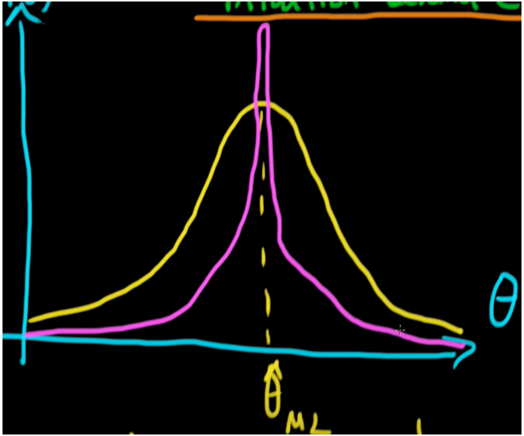
Ye kya hai ?!

1. Asymptotic variance of the log likelihood estimate
2. Sensitivity of the parameter θ

How !? - Out of syllabus

Sensitivity of θ

$$\mathcal{I}_\theta = \mathbb{E}_x\left[\nabla_\theta \log p(x|\theta)\, \nabla_\theta \log p(x|\theta)^T\right] \equiv -\mathbb{E}_x\left[\nabla^2_\theta \log p(x|\theta)\right]$$



Asymptotic Variance

Once upon a time, there was the Law of Large Numbers: $P(|\bar{X} - \mathbb{E}(X)| > \epsilon) \to 0$.

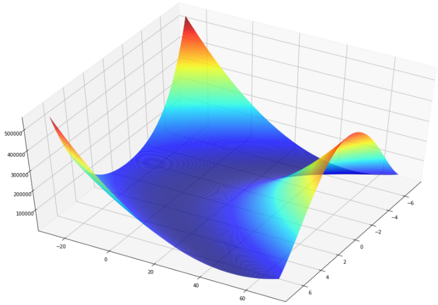
The Central Limit Theorem defines the rate of convergence: $P(\bar{X} - \mathbb{E}(X)) \to \mathcal{N}(0, \frac{\sigma^2}{n})$

Theorem (Asymptotic Variance of the maximum likelihood estimate)

$$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, \mathcal{I}_{\theta_0}^{-1})$$

Applications - Natural Gradient

So why is the Hessian important in optimization?
Why do we not use the Hessian in our optimization process?

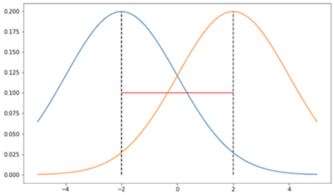


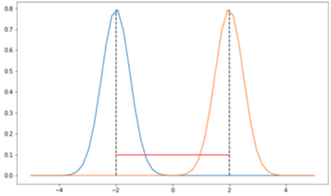
Theorem (Equivalency between Fisher and Hessian)

$$\mathcal{I}_\theta = -\mathbb{E}_{p(x|\theta)}[H_{log\, p(x|\theta)}]$$

Applications - Natural Gradient

So what is Gradient Descent? $\dfrac{-\nabla_\theta \mathcal{L}(\theta)}{||\nabla_\theta \mathcal{L}(\theta)||} = \lim_{\epsilon \to 0} \dfrac{1}{d} \arg\min_{d < |\epsilon|} \mathcal{L}(\theta + d)$





Parameter space or Distribution space?

Applications - Natural Gradient

Distribution space: KL Divergence!

Equivalence between KL divergence and Fisher information

$$KL[p(x|\theta)||p(x|\theta + d)] \approx \frac{1}{2} d^T \mathcal{I}_\theta d$$

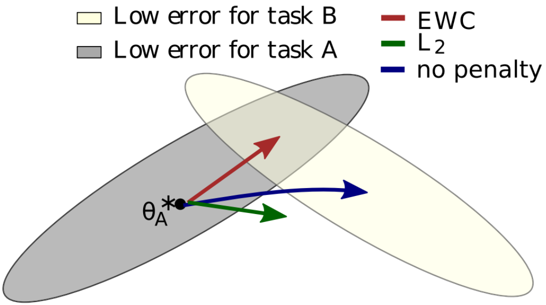
So a step in the parametric space is replaced by a step in the distribution space!
$\lim_{\epsilon \to 0} \arg\min_{d < |\epsilon|} \mathcal{L}(\theta + d) \implies \lim_{\epsilon \to 0} \arg\min_{d : KL[p_\theta || p_{\theta+d}] = \epsilon} \mathcal{L}(\theta + d)$

Theorem (Natural Gradient)

$$\hat{\nabla}_\theta \mathcal{L}(\theta) = \mathcal{I}_\theta^{-1} \nabla_\theta \mathcal{L}(\theta)$$

So why is Natural Gradient not popular? Ummm, Fisher matrix and inversion is expensive? So can we approximate the Fisher matrix? Yes, as done in *Adam* optimizer!

Applications - Elastic Weight Consolidation

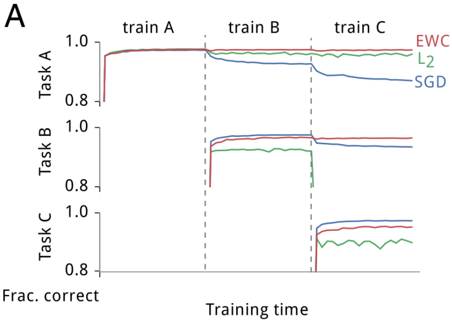


Extending $\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) - \log p(\mathcal{D})$ to tasks $\mathcal{D}_A, \mathcal{D}_B$:
$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}_B|\theta) + \log p(\theta|\mathcal{D}_A) - \log p(\mathcal{D}_B)$

Applications - Elastic Weight Consolidation

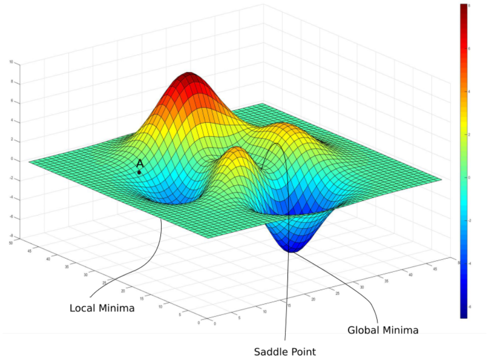
Theorem (Elastic Weight Consolidation)

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2}\mathcal{I}_i(\theta - \theta^*_{A,i})^2$$



"Overcoming catastrophic forgetting in neural networks", PNAS

Applications - Expected Gradient Length



Expected Gradient Length - Classification

$$x^*_{EGL} = \arg\max_x \sum_i P(y_i|x;\theta)\,||\,\nabla_\theta l\,(\mathcal{L} \cup (x, y_i);\theta)\,||$$

Applications - Expected Gradient Length

Can we derive this result?

Expected Gradient Length - Classification

$$x^*_{EGL} = \arg\max_x \sum_i P(y_i|x;\theta) \, || \, \nabla_\theta l \, (\mathcal{L} \cup (x, y_i); \theta) \, ||$$

Fisher to the rescue!

Applications - Expected Gradient Length

Asymptotic Variance

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow \mathcal{N}(0, \mathcal{I}_{\theta_0}^{-1})$$

Minimizing the variance is same as maximizing the Fisher Information

$$\max_q \int q(y|x,\theta) \, ||\nabla_\theta l(x,y,\theta)||^2 \, dy \, dx$$

Maximizing q is same as selecting unlabelled x having largest gradient

$$x^*_{EGL} = \arg\max_x \sum_i q(y_i|x;\theta) \, || \, \nabla_\theta l \, (x, y_i, \theta) \, ||^2$$

"Active Learning for Speech Recognition: the Power of Gradients", NIPS Workshop 2016

Thank You!

Thank you!