Natural Gradient Descent

Yutao Chen

⧖ Oct 03 2025
⧗ Oct 11 2025

Contents

Natural gradient descent (NGD) (Amari, 1998) is a second-order optimization method for parametrized probability distributions $p(\boldsymbol{x}|\boldsymbol{\theta})$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta F^{-1}(\boldsymbol{\theta}_t)\nabla J(\boldsymbol{\theta}_t),$$

where $F^{-1}(\boldsymbol{\theta}_t)$ is the inverse of the *Fisher information matrix* (FIM).

Definition 1 (Fisher Information Matrix)

The Fisher information matrix $F(\boldsymbol{\theta})$ for $p(x|\boldsymbol{\theta})$ is defined as the variance of the *score* function $\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta})$[1]:

$$F(\boldsymbol{\theta}) = \mathbb{E}_{p(x|\boldsymbol{\theta})}\left[(\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}))(\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}))^\top\right],$$

or equivalently the negative expected Hessian of the log likelihood

$$F(\boldsymbol{\theta}) = -\mathbb{E}_{p(x|\boldsymbol{\theta})}\left[\nabla^2_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta})\right].$$

Kullback-Leibler Divergence

Recall that [[Gradient Descent]] can be derived from

$$\arg\min_{\boldsymbol{\theta}} \; J(\boldsymbol{\theta}_t) + \nabla J(\boldsymbol{\theta}_t)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_t) + (2\eta)^{-1} \; \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_2^2,$$

which penalizes large updates by measuring the (squared) Euclidean distance $\|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_2^2$ between parameter.

[1] The expectation of the score function $\mathbb{E}_{p(x|\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta})] = 0$.

For probabilistic models, however, the parameter space is generally *not* Euclidean. Instead, a more natural penalty would be measuring the Kullback-Leibler (KL) divergence between the induced distributions:

$$\arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t) + \nabla J(\boldsymbol{\theta}_t)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_t) + \eta^{-1} \mathbb{D}_{\mathrm{KL}}(p(x|\boldsymbol{\theta}_t) \parallel p(x|\boldsymbol{\theta})).$$

Unfortunately, $\mathbb{D}_{\mathrm{KL}}(p\|q)$ do not have an analytical form in general. We instead consider its second-order Taylor approximation. Let $f(\boldsymbol{\theta}) = \mathbb{D}_{\mathrm{KL}}(p(x|\boldsymbol{\theta}_t) \parallel p(x|\boldsymbol{\theta}))$ and $\delta\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}_t$, and we have

$$f(\boldsymbol{\theta}) \approx f(\boldsymbol{\theta}_t) + \nabla f(\boldsymbol{\theta}_t)^T \delta\boldsymbol{\theta} + \frac{1}{2} \delta\boldsymbol{\theta}^\top \nabla^2 f(\boldsymbol{\theta}_t) \delta\boldsymbol{\theta}.$$

1. The first term is trivially zero as
$$f(\boldsymbol{\theta}_t) = \mathbb{D}_{\mathrm{KL}}(p(x|\boldsymbol{\theta}_t) \parallel p(x|\boldsymbol{\theta}_t)) = 0.$$
2. The second term is also zero as
$$\begin{aligned} \nabla f(\boldsymbol{\theta}_t) &= -\mathbb{E}_{p(x|\boldsymbol{\theta}_t)} \left[\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}) \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \right] \\ &= -\int \frac{p(x|\boldsymbol{\theta}_t)}{p(x|\boldsymbol{\theta}_t)} \nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta}_t) \, \mathrm{d}x = 0, \end{aligned}$$
where ∇ and $\int$ are exchangeable by the *Leibniz integral rule*.
3. The third term is non-zero. However, note that by Definition 1
$$\begin{aligned} \nabla^2 f(\boldsymbol{\theta}_t) &= -\mathbb{E}_{p(x|\boldsymbol{\theta}_t)} \left[\nabla^2_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}) \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \right] \\ &= -\mathbb{E}_{p(x|\boldsymbol{\theta}_t)} [\nabla^2_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}_t)] = F(\boldsymbol{\theta}_t). \end{aligned}$$

Therefore, we have $\mathbb{D}_{\mathrm{KL}}(p(x|\boldsymbol{\theta}_t) \parallel p(x|\boldsymbol{\theta})) \approx \frac{1}{2} \delta\boldsymbol{\theta}^\top F(\boldsymbol{\theta}_t) \delta\boldsymbol{\theta}$. Plugging back into the optimization objective, we have

$$\arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t) + \nabla J(\boldsymbol{\theta}_t)^\top \delta\boldsymbol{\theta} + (2\eta)^{-1} \delta\boldsymbol{\theta}^\top F(\boldsymbol{\theta}_t) \delta\boldsymbol{\theta},$$

which can be solved by $\boldsymbol{\theta}^* = \boldsymbol{\theta}_t - \eta F^{-1}(\boldsymbol{\theta}_t) \nabla J(\boldsymbol{\theta}_t)$.

Efficiency and Approximations

A major drawback of NGD is that computing and inverting the FIM is expensive. Therefore, efficient approximation methods or alternative routines have been of particular research interests.

Empirical Fisher

Recall that FIM can be defined as the variance of the score function

$$F(\boldsymbol{\theta}) = \mathbb{E}_{p(x|\boldsymbol{\theta})}\left[(\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}))(\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}))^\top\right].$$

Empirical Fisher proposes to approximate FIM by replacing $p(x|\boldsymbol{\theta})$ with the empirical distribution $p_{\mathcal{D}}(x)$, where $\mathcal{D}$ is a dataset:

$$F(\boldsymbol{\theta}) \approx \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} (\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}))(\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}))^\top.$$

Exponential Family Distributions

For an exponential family distribution with natural parameters $\boldsymbol{\lambda}$ and corresponding moment parameters $\boldsymbol{\mu}$, one can show that

$$F(\boldsymbol{\lambda})^{-1} \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}).$$

That is, the natural gradient w.r.t $\boldsymbol{\lambda}$ equals the regular gradient w.r.t. $\boldsymbol{\mu}$.

This result, under certain circumstances, conveniently allows performing NGD without actually computing FIM. We refer the readers to Khan & Rue (2023) and Murphy (2023) Section 6.4.5 for more context.

References

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*(2), 251–276. https://doi.org/10.1162/089976698300017746

Khan, M. E., & Rue, H. (2023). The Bayesian Learning Rule. *Journal of Machine Learning Research*, *24*(281), 1–46. http://jmlr.org/papers/v24/22-0291.html

Murphy, K. P. (2023). *Probabilistic Machine Learning: Advanced Topics*. MIT Press. http://probml.github.io/book2