The Intrinsicness of Gradient Methods

Jingfeng Wu

Created: April 2018
Last updated: June 27, 2020

We optimize a loss in the form of

$$L(\theta) = \mathbb{E}_z[\ell(z;\theta)]. \tag{1}$$

Typically, $z = (x, y)$ is the data pair. The expectation is taken over some distribution over z, which could be the population distribution or the empirical distribution.

Fixing initialization, by running an (convergent) optimization method we obtain a path from the initial point to the accumulation point. We call it optimization path. Then a natural question arises from the viewpoint of Differential Geometry:

For changing of parameterization, will a certain optimization method generate the same optimization path?

If the answer is yes, we say the optimization method is intrinsic (with respect to the reparameterization). In the following, we will show that NGD is intrinsic for any invertible regular reparameterization, while GD is intrinsic for only orthogonal regular reparameterization.

1 The Intrinsicness of Gradient Descent

Theorem 1 (GD is intrinsic for orthogonal transformations)**.** *NGD is invariant for any **orthogonal**, smooth reparameterization $\theta = \zeta(\gamma)$.*

Proof. For parameterization γ and $\theta = \zeta(\gamma)$, by assumption we have $J_\zeta^T J_\zeta = J_\zeta J_\zeta^T = I$. GD updates as

$$\Delta\,\theta = -\eta\,\nabla_\theta\,L(\theta), \quad \Delta\,\gamma = -\eta\,\nabla_\gamma\,L(\zeta(\gamma)). \tag{2}$$

To see the intrinsicness of GD, we only need to show that $\zeta(\gamma+\Delta\,\gamma)) \approx \theta+\Delta\,\theta$. Suppose the step size η is small enough for applying Taylor's expansion, then $\zeta(\gamma+\Delta\,\gamma) \approx \zeta(\gamma)+J_\zeta\cdot\Delta\,\gamma = \theta+J_\zeta\cdot\Delta\,\gamma$, where J_ζ is the Jacobian of transformation ζ. Thus we remind to show

$$\Delta\,\theta = J_\zeta\,\Delta\,\gamma. \tag{3}$$

Via chain rule, one have $\nabla_\gamma\,L(\zeta(\gamma)) = J_\zeta^T\,\nabla_\theta\,L(\theta)$, thus

$$J_\zeta\,\Delta\,\gamma = -J_\zeta\eta J_\zeta^T\,\nabla_\theta\,L(\theta) = -\eta\,\nabla_\theta\,L(\theta) = \Delta\,\theta. \tag{4}$$

The proof is completed. □

Remark: One can check that GD is not invariant for general **one-one** transformations.

2 The Intrinsicness of Natural Gradient Descent

In this section we focus on the invariance of gradient methods e.g., natural gradient descent (NGD) and gradient descent (GD), with respect to reparameteraization. Thus we do not rigorously distinguish Fisher, Empirical Fisher, Gradient Second Moment or Generalized Gauss-Newton matrix [1, 3]. Our discussion holds for all of them.

Definition 1 (Fisher Information)**.** For loss (1), the Fisher Information is defined as

$$F(\theta) = \mathbb{E}_z[\nabla_\theta\,\ell(z;\theta)\,\nabla_\theta\,\ell(z;\theta)^T]. \tag{5}$$

Definition 2 (Natural Gradient Descent)**.** For loss (1), the natural gradient descent (NGD) updates as [3]

$$\theta_{t+1} = \theta_t - \eta F(\theta_t)^{-1} \nabla_\theta L(\theta_t). \tag{6}$$

Theorem 2 indicates that NGD is invariant for any one-one (and regular) reparameterization.

Theorem 2 (NGD is intrinsic for invertible transformations)**.** *NGD is invariant for any* ***one-one****, smooth reparameterization* $\theta = \zeta(\gamma)$*.*

Proof. For parameterization γ and $\theta = \zeta(\gamma)$, NGD updates as

$$\Delta\theta = -\eta F(\theta)^{-1} \nabla_\theta L(\theta), \quad \Delta\gamma = -\eta F(\zeta(\gamma)))^{-1} \nabla_\gamma L(\zeta(\gamma)). \tag{7}$$

To see the intrinsicness of NGD, we only need to show that $\zeta(\gamma+\Delta\gamma)) \approx \theta+\Delta\theta$. Suppose the step size η is small enough for applying Taylor's expansion, then $\zeta(\gamma+\Delta\gamma) \approx \zeta(\gamma)+J_\zeta \cdot \Delta\gamma = \theta+J_\zeta \cdot \Delta\gamma$, where J_ζ is the Jacobian of transformation ζ. Thus we remind to show

$$\Delta\theta = J_\zeta \, \Delta\gamma. \tag{8}$$

Via chain rule, one have $\nabla_\gamma L(\zeta(\gamma)) = J_\zeta^T \nabla_\theta L(\theta)$ and

$$F(\zeta(\gamma)) = \mathbb{E}_z[\nabla_\gamma \ell(z;\zeta(\gamma)) \nabla_\gamma \ell(z;\zeta(\gamma))^T] = J_\zeta^T \mathbb{E}_z[\nabla_\theta \ell(z;\theta) \nabla_\theta \ell(z;\theta)^T] J_\zeta = J_\zeta^T F(\theta) J_\zeta. \tag{9}$$

Thus

$$\Delta\gamma = -\eta F(\zeta(\gamma))^{-1} \nabla_\gamma L(\zeta(\gamma)) = -\eta J_\zeta^{-1} F(\theta)^{-1} J_\zeta^{-T} J_\zeta^T \nabla_\theta L(\theta) = -\eta J_\zeta^{-1} F(\theta)^{-1} \nabla_\theta h(\theta), \tag{10}$$

and

$$J_\zeta \cdot \Delta\gamma = -J_\zeta \cdot \eta J_\zeta^{-1} F(\theta)^{-1} \nabla_\theta L(\theta) = -\eta F(\theta)^{-1} \nabla_\theta L(\theta) = \Delta\theta. \tag{11}$$

Hence the proof is finished. □

Remark: See [2] for discussions about overparameterized transformations.

3 The Intrinsicness of Newton's Method

Definition 3. For loss (1), the Newton's method updates as

$$\theta_{t+1} = \theta_t - \eta \, (H(\theta_t))^{-1} \nabla_\theta L(\theta_t), \tag{12}$$

where $H(\theta_t) = \nabla_\theta^2 L(\theta_t)$ is the Hessian.

Different from NGD, Newtons' method uses $H(\theta)$ as preconditioning instead of $F(\theta)$. For $\theta = \zeta(\gamma)$, we have

$$H(\zeta(\gamma)) = \nabla_\gamma^2 L(\zeta(\gamma)) = \nabla_\gamma^2 \zeta(\gamma) \cdot \nabla_\theta L(\theta) + J_\zeta^T \nabla_\theta^2 L(\theta) J_\zeta = \nabla_\gamma^2 \zeta(\gamma) \cdot \nabla_\theta L(\theta) + J_\zeta^T H(\theta) J_\zeta. \tag{13}$$

Comparing with the proof of Theorem 2 and the fact that

$$F(\zeta(\gamma)) = J_\zeta^T F(\theta) J_\zeta, \tag{14}$$

we know that Newton's method is intrinsic if and only if

$$\nabla_\gamma^2 \zeta(\gamma) \cdot \nabla_\theta L(\theta) = 0 \quad \Leftrightarrow \quad \nabla_\gamma^2 \zeta(\gamma) = 0, \tag{15}$$

that is, $\theta = \zeta(\gamma)$ is an affine transformation.

Theorem 3 (Newton's method is intrinsic for affine transformations)**.** *NGD is invariant for any* ***affine****, smooth reparameterization* $\theta = \zeta(\gamma)$*.*

References

[1] Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation. *arXiv preprint arXiv:1905.12558*, 2019.

[2] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. *arXiv preprint arXiv:1711.01530*, 2017.

[3] James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.