Parametrization-invariant gradient methods

Contents

Reminder

Suppose we are given $f : \Theta \to \mathbb{R}$ to optimize (a typical example in machine learning is : $f : \theta \mapsto \sum_i \log(p_\theta(x_i))$).

We recall that a gradient method is given by

$$\theta \leftarrow \theta - \eta \frac{\partial f}{\partial \theta}.^{1} \tag{1}$$

However, the behavior of (1) depends on the parametrization, and as we saw in the last talk, for a given problem, *any* point in the half-space $\frac{\partial f}{\partial \theta} > 0$ is the endpoint of a gradient step for some given parametrization.

Let us now consider a gradient step in some parametrization θ_1, and rewrite *this* gradient step in another parametrization θ_2.

- If θ_2 is given by an affine transform, $\theta_2 = A\theta_1$, (1) becomes

$$\theta_2 \leftarrow \theta_2 - \eta M^{-1} \frac{\partial f}{\partial \theta_2}, \tag{2}$$

  with $M = A^T A$.

- With more general transformations $\theta_2 = g(\theta_1)$, we find that the matrix M depends on θ.

$$\theta_2 \leftarrow \theta_2 - \eta M^{-1}(\theta_2) \frac{\partial f}{\partial \theta_2}, \tag{3}$$

  with $M(\theta_2) = (\frac{\partial g}{\partial \theta_1})^T (\frac{\partial g}{\partial \theta_1})$, with θ_1 the preimage of θ_2 under g.

[1] Until Section 1.3, we use the convention that $\frac{\partial f}{\partial \theta}$ is the column vector satisfying $(\frac{\partial f}{\partial \theta})_i = \frac{\partial f}{\partial \theta_i}$ (instead of being a row vector, as we explained it should be in the previous talk).

Remark. It could be argued that the η in the gradient updates is unnecessary, since it is redundant with M. However, η has a concrete meaning: it is the time step of the time discretization corresponding to the gradient method (we recall the continuous gradient descent: $\frac{\mathrm{d}\theta}{\mathrm{d}t} = M^{-1}(\theta)\frac{\partial f}{\partial \theta}$).

As we see from equation (3), it would be interesting to be able to define a $M(\theta)$ such that the gradient step $M^{-1}(\theta)\frac{\partial f}{\partial \theta}$ does not depend on the parametrization. We are going to see three different ways to do so.

1 Invariant gradient descents

The three possibilities we are going to study are the Newton method, the outer product metric, and the natural metric (of which the Fisher metric, which will be studied in the next talk, is a particular case).

1.1 Newton method

To find the minimum of f, we try to solve $f'(\theta) = 0$ with the Newton method. This yields:

$$\theta \leftarrow \theta - (\mathrm{Hess}(f))^{-1}\frac{\partial f}{\partial \theta}, \tag{4}$$

which is a particular form of the general gradient descent $\theta \leftarrow \theta - \eta M^{-1}(\theta)\frac{\partial f}{\partial \theta}$, with $M = \mathrm{Hess}(f)$.

It is however important to notice that while equation (4) is independent with respect to affine reparametrization, it is not independent with respect to non-linear reparametrization.

Example. Suppose we are optimizing the function $f : x \mapsto x^2$ with the Newton method. We jump in one step to the minimum. If we set $x = \log y$ and work with y instead, however, we need to optimize $\tilde{f} : y \mapsto (\log y)^2$, for which the Newton method does not work well (there is an inflexion point, and the function is concave near $+\infty$).

In general, we can notice that Newton method is equivalent to looking at the second order Taylor expansion of f, and jump to its minimum, which explains the three following properties of the Newton method:

- The Newton method is invariant with respect to affine transformations.
- The Newton method is *not* invariant with respect to non linear transformations.
- Near a minimum, the second order Taylor expansion *is* well-defined, and consequently, the Newton method is fully invariant with respect to any reparametrization.

Consequently, it is better not to use the Newton method far away from a minimum.

To obtain full invariance[2] with respect to reparametrization of θ, we start by recalling another expression of gradient descents.

[2]This must be read "full invariance in continuous time", i.e., when $\eta \to 0$: two different

1.2 Gradient descent as penalized minimization

As we showed during the last talk, the general gradient step $\theta \leftarrow \theta - \eta M^{-1}(\theta)\frac{\partial f}{\partial \theta}$ can be rewritten as the following penalized optimization problem.

$$\theta^{t+1} = \text{argmin}_\theta\{f(\theta) + \frac{1}{2\eta}\|\theta - \theta^t\|^2_{\theta^t}\}, \tag{5}$$

at first order in η, where $\|x\|_{\theta^t} := x^T M^{-1}(\theta^t)x$ is the norm defined by M.[3]

Consequently, instead of trying to define M directly, we are going to find a reasonable penalty for our problem, and use it to define the metric. This will yield the outer product metric and the natural metric.

1.3 Outer product metric

This section applies only if f has a "natural" decomposition: $f = \frac{1}{N}\sum_i f_i$, with $N \geqslant \dim\Theta$. This situation often arises in machine learning.

We start by defining the outer product metric.

We are given $f : \Theta \to \mathbb{R}$.

A change $\theta \leftarrow \theta + \delta\theta$ induces a change on f: $f \leftarrow f + \delta f$. It therefore seems reasonable to set:

$$\|\delta\theta\|^2 := \|\delta f\|^2. \tag{6}$$

We can compute the corresponding metric: we have $\delta f = \frac{\partial f}{\partial \theta}\delta\theta$, so $\|\delta f\|^2 = (\frac{\partial f}{\partial \theta}\delta\theta)^T \frac{\partial f}{\partial \theta}\delta\theta = \delta\theta^T(\frac{\partial f}{\partial \theta})^T \frac{\partial f}{\partial \theta}\delta\theta$.

The gradient descent we would like to write is therefore $\theta \leftarrow \theta - \eta M^{-1}(\theta)\frac{\partial f}{\partial \theta}$, with $M = (\frac{\partial f}{\partial \theta})^T \frac{\partial f}{\partial \theta}$, but M is a rank-1 matrix, and is consequently not invertible. Geometrically speaking, the cause for this is that moving in the direction of the level sets is not penalized.

We therefore modify the previous definition:

Definition 1. *Let $f : \Theta \to \mathbb{R}$, with $f = \frac{1}{N}\sum_{i=1}^n f_i$, $N \geqslant \dim\Theta$. The outer product norm of a change $\delta\theta$ is by definition:*

$$\|\delta\theta\|^2_{\text{OP}} := \frac{1}{N}\sum_{i=1}^N \|\delta f_i\|^2. \tag{7}$$

The corresponding metric is[4]

$$M(\theta) = \frac{1}{N}\sum_{i=1}^N (\frac{\partial f_i}{\partial \theta})^T \frac{\partial f_i}{\partial \theta}. \tag{8}$$

Now, each i contributes to the metric by adding a rank one matrix. Consequently, it is necessary to have $N \geqslant \dim\Theta$ for M to be invertible. This is reasonable: at least N observations are needed in order to estimate N parameters.

We also give a characterization of the outer product gradient step.

"invariant" algorithms will coincide up to second order in η, and consequently, they will be approximations of the same trajectory. This is not the case for non invariant algorithms.

[3] We are now using the convention that $\frac{\partial f}{\partial \theta}$ is a row vector.

[4] The proof is the same as above.

Proposition 2. *For a given total increase δf, the $\delta\theta$ selected by the outer product metric is the only minimizer the variance of the δf_i.*

Proof. Let us fix $\|\delta f\|^2$.

We have $\mathrm{Var}(\delta f_i) = \frac{1}{N}\sum_i \|\delta f_i\|^2 - \|\frac{1}{N}\sum_i \delta f_i\|^2 = \frac{1}{N}\sum_i \|\delta f_i\|^2 - \|\delta f\|^2$. Consequently,

$$\mathrm{Var}(\delta f_i) = \|\delta\theta\|^2_{\mathrm{OP}} - \|\delta f\|^2. \tag{9}$$

Since $\|\delta f\|$ is a constant, the proposition follows from equation (5).

□

The main advantage of this metric is that it is easy to compute if we know the $\frac{\partial f_i}{\partial \theta}$, which is the case with neural networks for example.

However, this metric also has shortcomings. For example, in dimension 1, with $N = 1$, we find the following gradient update:

$$\theta \leftarrow \theta - \eta \frac{f'(\theta)}{(f'(\theta))^2} = \theta - \eta \frac{1}{f'(\theta)}, \tag{10}$$

which is ill-behaved when θ is close to the minimum. This can be solved by taking $N \geqslant \dim\Theta + 1$ (this solution works only if some functions in the decomposition have a non-zero derivative at θ). Moreover, step-size control is needed. Another problem is that the gradient update with the metric (8) it is not independent with respect to reparametrization of f (this can be partly solved by using a homogeneous η).

Example. Consider the function $f : x \mapsto \frac{1}{2}x^2$. As we have seen, the outer product metric is ill-behaved. But now, if we set $f_1 : x \mapsto x^2 - x$ and $f_2 : x \mapsto x$, the update is given by $x \leftarrow x - \eta\frac{2x}{(2x-1)^2+1}$. Now, the update around 0 is fine, but if we start far away from it, the gradient step is roughly equal to $\eta\frac{1}{2x}$, so many steps will be needed to reach the optimum: the algorithm essentially tries to improve the objective by 1 at each step.

Let us now study the natural metric on θ.

1.4 Natural metric on Θ

Suppose we are trying to minimize f such that there exists a set Y and two maps $L : Y^2 \to \mathbb{R}$, and $y : \Theta \to Y$ such that

$$f(\theta) = L(y(\theta), \bar{y}) \tag{11}$$

This appears in prediction problems: $y(\theta)$ is the prediction, $\bar{y}$ is the observation, and L is the loss, quadratic for example.

If we have a metric $\|\delta y\|^2$ on Y, we can turn it into a metric on Θ, since a change $\theta \leftarrow \theta + \delta\theta$ induces a change $y \leftarrow y + \delta y$.

Definition 3. *Let $\|.\|_Y$ be a norm on y. We define the natural metric on Θ:*

$$\|\delta\theta\|^2_{\mathrm{nat}} := \|\delta y\|^2_Y, \tag{12}$$

where δy is the change in y induced by the change $\theta \leftarrow \theta + \delta\theta$.

This metric can be computed easily. Indeed, if $\|y\|_Y^2 = y^T M_Y y$, we have $\|\delta\theta\|_{\text{nat}}^2 = \|\frac{\partial y}{\partial \theta}\delta\theta\|_Y^2 = \delta\theta^T \frac{\partial y}{\partial \theta}^T M_Y \frac{\partial y}{\partial \theta}\delta\theta$. In other words, we find

$$M_\theta = \frac{\partial y}{\partial \theta}^T M_Y \frac{\partial y}{\partial \theta}. \tag{13}$$

A prediction can often be a probability distribution on the variable we are interested in. For example, if we are trying to predict the weather tomorrow, a prediction could be: "it will rain with probability p, and it will be sunny with probability $1-p$".

In that case, Y is itself a *family of probability distributions* on some set Z we are interested in (in the weather case, $Z = \{\text{rain}, \text{no rain}\}$, and Y is the set of Bernoulli distributions on Z). Then, a reasonable loss is $L = -\log y(\bar{z})$, for some $\bar{z} \in Z$ (in our example $\bar{z}$ would be the observed weather).

Notice for example that if $y = \mathcal{N}(m, \sigma^2)$, we find the square loss.

We now have to find a metric on Y. If Y is a family of probability distributions, an interesting choice is the Fisher metric, which will be discussed in the next talk.