Gradient Descent (and beyond)

Kun He (何琨)

Data Mining and Machine Learning Lab
(John Hopcroft Lab)
Huazhong University of Science & Technology

brooklet60@hust.edu.cn

2022年5月

Table of contents

Table of Contents

1. Introduction
   - Review
   - Local minimizer
2. Taylor Expansions
   - Taylor expansions
   - First and second order Taylor approximations
3. Search Direction Methods
4. Gradient Descent
5. Adagrad
6. Newton's method
7. A Simple Example
8. Best Practices

Review

- In the previous lecture on Logistic Regression we wrote down expressions for the parameters in our model as solutions to optimization problems that do not have closed form solutions.
- Specifically, given data $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$ we saw that

  $$\hat{w}_{\text{MLE}} = \operatorname*{arg\,min}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \sum_{i=1}^n \log(1 + e^{-y_i(w^T x_i + b)})$$

  and

  $$\hat{w}_{\text{MAP}} = \operatorname*{arg\,min}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \sum_{i=1}^n \log(1 + e^{-y_i(w^T x_i + b)}) + \lambda w^T w.$$

- These notes will discuss general strategies to solve these problems and, therefore, we abstract our problem to

  $$\min_w \ell(w)$$

  where $\ell \colon \mathbb{R}^d \to \mathbb{R}$.

Review

While we will discuss some basic algorithmic strategies here.

We want to minimize a **convex**, **continuous** and **differentiable** loss function $\ell(w)$.

- ℓ is convex. This allows us to assert that any local minimum we find is also a global minimum, and helps simplify our discussion of Newton’ s method.
- ℓ is at least thrice continuously differentiable. We are going to extensively use Taylor approximations and this assumption simplifies the discussion greatly.
- There are no constraints placed on w. Adding constraints is a level of complexity we will not address here.

In this section we discuss two of the most popular "hill-climbing" algorithms, gradient descent and Newton's method for $\min_w \ell(w)$.

What is a (local) minimizer

- The first question we might ask is what it actually means to solve $\min_w \ell(w)$. We call w^* a local minimizer of ℓ if:

Local minimizer:

There is some $\epsilon > 0$ such that for $\{w \mid \|w - w^*\|_2 < \epsilon\}$ we have that $\ell(w^*) \leq \ell(w)$.

- Our earlier assumption that ℓ is convex implies that if we find such a w^* we can immediately assert that $\ell(w^*) \leq \ell(w)$ for all $w \in \mathbb{R}^d$.
- we can also define a strict local minimizer by forcing $\ell(w^*) < \ell(w)$.
- Notably, some convex functions have no strict local minimizers, e.g., the constant function $\ell(w) = 1$ is convex.
- Some convex functions have no finite local minimizers, e.g., for any non-zero vector $c \in \mathbb{R}^d$, $c^T w$ is convex but can be made arbitrarily small by letting appropriately.
- A key necessary condition for a point to be a local minimizer is that the gradient of ℓ is zero at w^*, i.e., $\nabla \ell(w^*) = 0$.
- Assuming the gradient at w^* is zero, a sufficient condition for the point to be a strict local minimizer is for the Hessian, i.e., $\nabla^2 \ell(w^*)$ to be positive definite.

Table of Contents

1. Introduction
   - Review
   - Local minimizer
2. Taylor Expansions
   - Taylor expansions
   - First and second order Taylor approximations
3. Search Direction Methods
4. Gradient Descent
5. Adagrad
6. Newton's method
7. A Simple Example
8. Best Practices

Taylor expansions

- While we made some assumptions on ℓ they do not actually tell us much about the function—it could behave in all sorts of ways. Moreover, as motivated by the logistic regression example it may not be so easy to work with the function globally.
- Therefore, we will often leverage local information about the function ℓ. We accomplish this through the use of first and second order Taylor expansions.

Taylor expansions

First order Taylor expansion

The first order Taylor expansion of centered at can be written as

$$\ell(w+p) \approx \ell(w) + g(w)^T p,$$

where $g(w)$ is the gradient of ℓ evaluated at w i.e., $(g(w))_j = \frac{\partial \ell}{\partial w_j}(w)$ for $j = 1, \ldots, d$.

Second order Taylor expansion

The second order Taylor expansion of centered at ℓ can be written as

$$\ell(w+p) \approx \ell(w) + g(w)^T p + \frac{1}{2} p^T H(w) p,$$

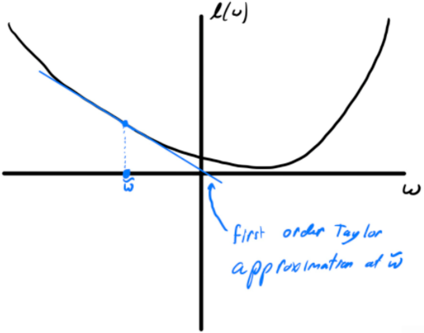
where $H(w)$ is the Hessian of ℓ evaluated at w, i.e.,

$$[H(w)]_{i,j} = \frac{\partial^2 \ell}{\partial w_i \partial w_j}(w)$$

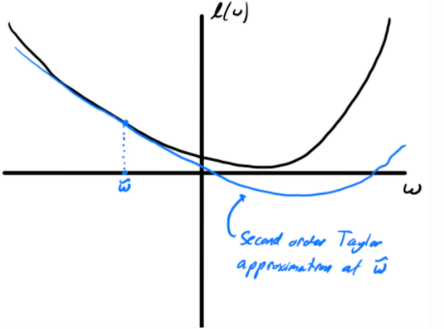
for $j = 1, \ldots, d$.

First and second order Taylor approximations

- These correspond to linear and quadratic approximations of ℓ.
- In general, these approximations are reasonably valid if p is small (concretely, first order has error $\mathcal{O}(\|p\|_2^2)$ and second order has error $\mathcal{O}(\|p\|_2^3)$).



$$\ell(w+p) \approx \ell(w) + g(w)^T p$$



$$\ell(w+p) \approx \ell(w) + g(w)^T p + \frac{1}{2} p^T H(w) p$$

Table of Contents

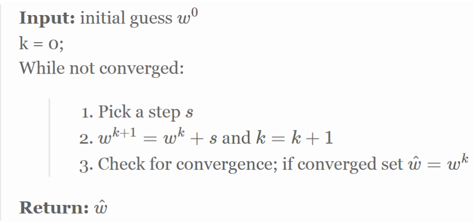
1. Introduction
   - Review
   - Local minimizer
2. Taylor Expansions
   - Taylor expansions
   - First and second order Taylor approximations
3. Search Direction Methods
4. Gradient Descent
5. Adagrad
6. Newton's method
7. A Simple Example
8. Best Practices

Search direction methods

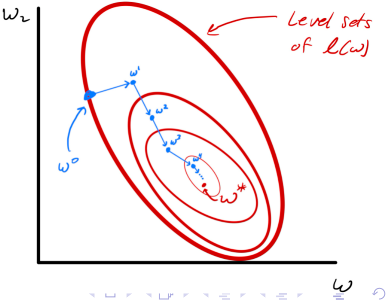
Loss function

$$\min_{w} \ell(w),$$

- The core idea is that given a starting point w^0 we construct a sequence of iterates $w^1, w^2, \ldots$ with the goal that $w^k \to w^*$ as $k \to \infty$.
- In a search direction method we will think of constructing w^{k+1} from w^k by writing it as $w^{k+1} = w^k + s$ for some "step" s.



```
Input: initial guess $w^0$
k = 0;
While not converged:
    1. Pick a step $s$
    2. $w^{k+1} = w^k + s$ and $k = k + 1$
    3. Check for convergence; if converged set $\hat{w} = w^k$
Return: $\hat{w}$
```



Search direction methods

Two key steps

There are two clearly ambiguous steps in the above algorithm:

- **How do we pick** s.
- how do we determine when we have converged.

We will spend most of our time addressing the former and then briefly touch on the latter —robustly determining convergence is one of the little details that a good optimization package should do well.

Table of Contents

1. Introduction
    - Review
    - Local minimizer
2. Taylor Expansions
    - Taylor expansions
    - First and second order Taylor approximations
3. Search Direction Methods
4. Gradient Descent
5. Adagrad
6. Newton's method
7. A Simple Example
8. Best Practices

Gradient descent

Core idea

Given we are currently at determine the direction in which the function decreases the fastest (at this point) and take a step in that direction.

Considering the linear approximation to at provided by the Taylor series

$$\ell(w^k + s) = \ell(w^k) + g(w^k)^T s$$

then the fastest direction to descend is simply $s \propto -g(w^k)$.

what we actually do in gradient descent is set s as

$$s = -\alpha g(w^k)$$

for some step size $\alpha > 0$.

Gradient descent

Correctness

There is always some small enough α such that

$$\ell(w^k - \alpha g(w^k)) < \ell(w^k).$$

Question: Why?

$$\ell(w^k + s) = \ell(w^k) + g(w^k)^T s$$

$$s = -\alpha g(w^k)$$

$$\alpha > 0$$

Gradient descent

Correctness

There is always some small enough α such that

$$\ell(w^k - \alpha g(w^k)) < \ell(w^k).$$

$$\ell(w^k - \alpha g(w^k)) = \ell(w^k) - \alpha g(w^k)^T g(w^k) + \mathcal{O}(\alpha^2).$$

Since

$$g(w^k)^T g(w^k) > 0$$

and $\alpha^2 \to 0$ faster than α as $\alpha \to 0$.

We conclude that for some sufficiently small $\alpha > 0$ we have that $\ell(w^k - \alpha g(w^k)) < \ell(w^k)$.

Determine the step size

- In classical optimization, α is often referred to as the step size (in this case $g(w^k)$ is the search direction).
- However, they can be more expensive then a fixed strategy for setting α. The catch is that setting α too small can lead to slow convergence and setting α too large can actually lead to divergence.

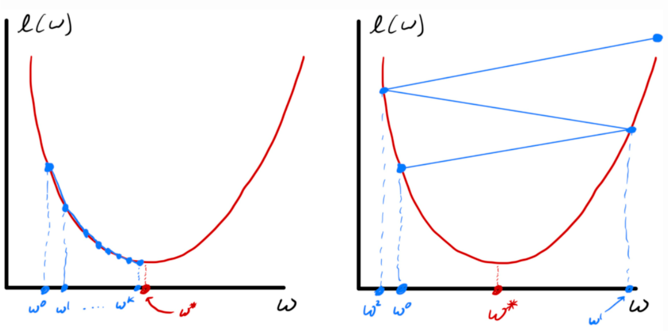


图: Choices of the step size that lead to convergence (left) or divergence (right).

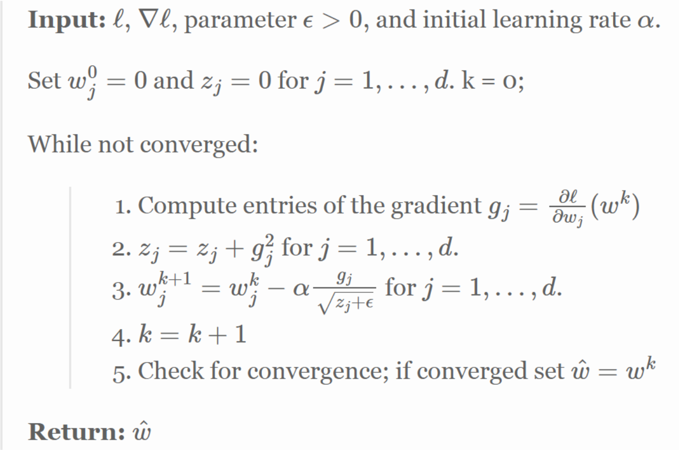
Table of Contents

1. Introduction
   - Review
   - Local minimizer
2. Taylor Expansions
   - Taylor expansions
   - First and second order Taylor approximations
3. Search Direction Methods
4. Gradient Descent
5. Adagrad
6. Newton's method
7. A Simple Example
8. Best Practices

Adagrad

- One option is to set the step-size adaptively for every feature.
- Adagrad accomplishes this by keeping a running average of the squared gradient with respect to each optimization variable.
- It then sets a small learning rate for variables with large gradients and a large learning rate for features with small gradients.
- This can be important if the entries of w are attached to features (e.g., in logistic regression we can associate each entry of w with a feature) that vary in scale or frequency.

Adagrad



```
Input: $\ell$, $\nabla\ell$, parameter $\epsilon > 0$, and initial learning rate $\alpha$.

Set $w_j^0 = 0$ and $z_j = 0$ for $j = 1, \dots, d$. k = 0;

While not converged:

    1. Compute entries of the gradient $g_j = \frac{\partial \ell}{\partial w_j}(w^k)$
    2. $z_j = z_j + g_j^2$ for $j = 1, \dots, d$.
    3. $w_j^{k+1} = w_j^k - \alpha \frac{g_j}{\sqrt{z_j + \epsilon}}$ for $j = 1, \dots, d$.
    4. $k = k + 1$
    5. Check for convergence; if converged set $\hat{w} = w^k$

Return: $\hat{w}$
```

Keypoint: Every dimension uses its own learning rate.

Question: Why add ϵ?

Table of Contents

1. Introduction
   - Review
   - Local minimizer
2. Taylor Expansions
   - Taylor expansions
   - First and second order Taylor approximations
3. Search Direction Methods
4. Gradient Descent
5. Adagrad
6. Newton's method
7. A Simple Example
8. Best Practices

Newton's method

Core idea

Use second order information (quadratic approximation).

$$\ell(w^k + s) \approx \ell(w^k) + g(w^k)^T s + \frac{1}{2} s^T H(w^k) s.$$

- We now chose a step by explicitly minimizing the quadratic approximation to ℓ at w^k.
- Recall that because ℓ is convex $H(w)$ is positive semi-definite for all w so this is a sensible thing to attempt.
- In fact, Newton's method has very good properties in the neighborhood of a strict local minimizer and once close enough to a solution it converges rapidly.

$$\boldsymbol{H}(\mathbf{w}) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial w_1^2} & \frac{\partial^2 \ell}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 \ell}{\partial w_1 \partial w_n} \\ \vdots & \cdots & \cdots & \vdots \\ \frac{\partial^2 \ell}{\partial w_n \partial w_1} & \cdots & \cdots & \frac{\partial^2 \ell}{\partial w_n^2} \end{pmatrix},$$

Newton's method

Core idea

Use second order information (quadratic approximation).

$$\ell(w^k + s) \approx \ell(w^k) + g(w^k)^T s + \frac{1}{2} s^T H(w^k) s.$$

- For simplicity, lets assume that $H(w^k)$ is positive definite.
- The gradient of our quadratic approximation is $g(w^k) + H(w^k)s$.
- This implies that s solves the linear system:

$$H(w^k)s = -g(w^k).$$

$$s = -H^{-1}(w^k)g(w^k).$$

Table of Contents

A simple example

- There is a simple example that clearly illustrates how incorporating second order information can help.
- Pretend the function was actually a strictly convex quadratic, i.e.,

$$\ell(w) = \frac{1}{2} w^T A w + b^T w + c$$

where A is a positive definite matrix, b is an arbitrary vector, and c is some number.

Question: How many steps does Newton converge?

A simple example

- Pretend the function was actually a strictly convex quadratic, i.e.,

$$\ell(w) = \frac{1}{2} w^T A w + b^T w + c$$

where A is a positive definite matrix, b is an arbitrary vector, and c is some number.

In this case, Newton converges in one step (since the strict global minimizer of w^* is the unique solution to $Aw = b$).

A simple example

- Pretend the function was actually a strictly convex quadratic, i.e.,

$$\ell(w) = \frac{1}{2} w^T A w + b^T w + c$$

where A is a positive definite matrix, b is an arbitrary vector, and c is some number.

Meanwhile, gradient descent yields the sequence of iterates

$$w^k = (I - \alpha A) w^{k-1} - \alpha b.$$

Using the fact that $w^* = (I - \alpha A) w^* - \alpha b$ we can see that

$$\begin{aligned} \|w^k - w^*\| &\leq \|I - \alpha A\|_2 \|w^{k-1} - w^*\|_2 && (1) \\ &\leq \|I - \alpha A\|_2^k \|w^0 - w^*\|_2. && (2) \end{aligned}$$

A simple example

- Therefore, as long as α is small enough such that all the eigenvalues of $I - \alpha A$ are in (-1,1), the iteration will converge—albeit slowly if we have eigenvalues close to ± 1.
- More generally, fig. 4 shows an example where we see the accelerated convergence of Newton’ s method as we approach the local minimizer.

The following figure shows an example where we see the accelerated convergence of Newton's method as we approach the local minimizer.

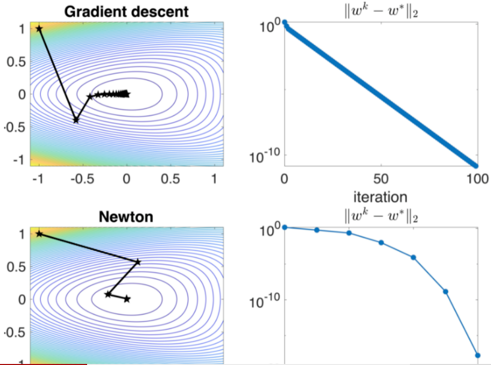


Table of Contents

1 Introduction
- Review
- Local minimizer

2 Taylor Expansions
- Taylor expansions
- First and second order Taylor approximations

3 Search Direction Methods

4 Gradient Descent

5 Adagrad

6 Newton's method

7 A Simple Example

8 Best Practices

Best Practices

- The matrix $H(w)$ scales $d \times d$ and is expensive to compute. A good approximation can be to **only compute its diagonal entries and multiply the update with a small step-size**. Essentially you are then doing a hybrid between Newton's method and gradient descent, where you weigh the step-size for each dimension by the inverse Hessian..
- To avoid divergence of Newton's method, a good approach is to **start with gradient descent (or even stochastic gradient descent) and then finish the optimization Newton's method**. Typically, the second order approximation, used by Newton's Method, is more likely to be appropriate near the optimum.

The End