Optimization for Data Science

F. Rinaldi[1]

1 Università degli Studi di Padova | Dipartimento Matematica

Padova
2020

Outline

Optimization for Data Science

Gradient Descent Schemes

- Gradient descent schemes traced back to Cauchy.
- Simplest way to minimize a differentiable function f on $\mathbb{R}^n$.
- We use a linear (first order) approximation of $f(x_k + d)$ to calculate the search direction at each iteration.
- First-order approximation:

$$f(x_k + d) = f(x_k) + \nabla f(x_k)^\top d + \beta_1(x_k, d),$$

with

$$\lim_{\|d\| \to 0} \frac{\beta_1(x_k, d)}{\|d\|} = 0.$$

Gradient Descent Schemes II

- In practice, we approximate $f(x_k + d)$ with the function $\eta_k(d)$ defined as follows:

$$\eta_k(d) := f(x_k) + \nabla f(x_k)^\top d.$$

- Then choose d_k as the direction such that:

$$\begin{aligned} &\min \eta_k(d) \\ &\|d\| = 1, \end{aligned}$$

- Equivalent to

$$\begin{aligned} &\min \nabla f(x_k)^\top d \\ &\|d\| = 1. \end{aligned}$$

Cauchy-Schwarz Inequality

- *Cauchy-Schwarz Inequality.* $|x^\top y| \le \|x\| \cdot \|y\|$.

Gradient Descent Schemes III

- Using the Cauchy-Schwarz inequality, we prove that the optimal direction is
$$d_k^* = -\nabla f(x_k)/\|\nabla f(x_k)\|.$$
- The classic *gradient method* calculate each iterate as follows:
$$x_{k+1} = x_k - \tilde{\alpha}_k \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}.$$
- By suitably redefining the stepsize
$$\alpha_k := \tilde{\alpha}_k/\|\nabla f(x_k)\|,$$
we then have
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

Gradient Method: Detailed Scheme

Algorithm 1 `Gradient method`

1 Choose a point $x_1 \in \mathbb{R}^n$
2 For $k = 1, \ldots$
3 If x_k satisfies some specific condition, then STOP
4 Set $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$, with $\alpha_k > 0$ a stepsize
5 End for

Lipschitz Continuous Gradient

Definition [Lipschitz Continuous Gradient]

A function $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz continuous gradient if there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n. \tag{1}$$

For functions with Lipschitz continuous gradient, we can prove the following result:

Proposition [LCG Inequality]

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function with Lipschitz continuous gradient. Then, for any $x, y \in \mathbb{R}^n$, we have

$$|f(x) - f(y) - \nabla f(y)^\top (x - y)| \leq \frac{L}{2}\|x - y\|^2. \tag{2}$$

The mean Theorem in Integral Form and Caucy-Schwarz Inequality

- *The mean theorem in integral form.* For a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we have, for all $d \in \mathbb{R}^n$, that the following equality holds:

$$f(z+d) - f(z) = \int_0^1 \nabla f(z+td)^\top d \, dt.$$

- *Cauchy-Schwarz Inequality.* $|x^\top y| \leq \|x\| \cdot \|y\|$.

Proof

Use the mean theorem in integral form, basic rule of integrals, Cauchy-Schwarz and finally use the fact that gradient is Lipschitz continuous:

$$
\begin{aligned}
& |f(x) - f(y) - \nabla f(y)^\top (x-y)| \\
= \quad & \left| \int_0^1 \nabla f(y + t(x-y))^\top (x-y) - \nabla f(y)^\top (x-y) \, dt \right| \\
& \text{(from basic rules of integrals)} \\
\leq \quad & \int_0^1 \left| \left(\nabla f(y + t(x-y)) - \nabla f(y)\right)^\top (x-y) \right| dt \\
& \text{(apply Cauchy-Schwarz)} \\
\leq \quad & \int_0^1 \|\nabla f(y + t(x-y)) - \nabla f(y)\| \cdot \|x-y\| \, dt \\
& \text{(gradient Lipschitz continuous)} \\
\leq \quad & \int_0^1 Lt\|x-y\|^2 dt = \frac{L}{2}\|x-y\|^2,
\end{aligned}
$$

Convex Case

- We have

$$f(x) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{L}{2}\|x - y\|^2. \tag{3}$$

Remark

- This inequality gives an upper bound over $f(x)$!
- Very useful to prove convergence results.

Equivalent Results

The following result prove some useful equivalence that we will be using in our proofs.

Proposition [Equivalence Results for LCG]

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. The following statements are equivalent:

(i) f has Lipschitz continuous gradient with constant $L > 0$;

(ii) $f(x) - f(y) - \nabla f(y)^\top (x - y) \leq \frac{L}{2}\|x - y\|^2$ for all $x, y \in \mathbb{R}^n$;

(iii) $f(x) \geq f(y) + \nabla f(y)^\top (x - y) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$ for all $x, y \in \mathbb{R}^n$;

(iv) $(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$ for all $x, y \in \mathbb{R}^n$.

A Useful Remark

Remark [Fixed Stepsize]

Let us consider gradient method with $\alpha_k = t > 0$. Using equation (3) (LCG convex case) where we set

$$x = x_{k+1} = x_k - \alpha_k \nabla f(x_k) = x_k - t\nabla f(x_k)$$

and

$$y = x_k,$$

we can write

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \nabla f(x_k)^\top (x_k - t\nabla f(x_k) - x_k) + \frac{L}{2}\|x_k - t\nabla f(x_k) - x_k\|^2 \\ &= -t\|\nabla f(x_k)\|^2 + \frac{Lt^2}{2}\|\nabla f(x_k)\|^2 = -(1 - \frac{Lt}{2})t\|\nabla f(x_k)\|^2. \end{aligned}$$

Since we want to get a stepsize that maximizes the reduction, we need to choose $\alpha_k = \frac{1}{L}$. Hence, we can write

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2. \tag{4}$$

Convergence Result for the Gradient Method

Theorem [Convergence of the Gradient Method]

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function with Lipschitz continuous gradient having Lipschitz constant $L > 0$. Gradient method with fixed stepsize $\alpha_k = 1/L$, satisfies:

$$f(x_{k+1}) - f(x^*) \leq \frac{2L\|x_1 - x^*\|^2}{k}.$$

$$\Downarrow$$

Remark

- We get **sublinear** rate!!
- We analyze $\mathcal{E}(x_{k+1})$ (Rate wrt $\mathcal{E}(x_k)$ by rescaling index k).

Proof of Convergence

Taking into account first order convexity conditions, we can write

$$f(x_k) \leq f(x^*) + \nabla f(x_k)^\top (x_k - x^*)$$

and thus we obtain, from Cauchy-Schwarz inequality the following:

$$f(x_k) - f(x^*) \leq \|\nabla f(x_k)\| \cdot \|x_k - x^*\|,$$

that is

$$\|\nabla f(x_k)\| \geq \frac{f(x_k) - f(x^*)}{\|x_k - x^*\|}. \tag{5}$$

By plugging (5) into

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2,$$

we have

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\left(\frac{f(x_k) - f(x^*)}{\|x_k - x^*\|}\right)^2. \tag{6}$$

Proof of Convergence II

Now, we prove that

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\|. \tag{7}$$

We simply use definition of x_{k+1} to get

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - t\nabla f(x_k) - x^*\|^2 \\
&= \|x_k - x^*\|^2 - \frac{2}{L}\nabla f(x_k)^\top (x_k - x^*) + \frac{1}{L^2}\|\nabla f(x_k)\|^2 \\
&\quad \text{(using } \left(\nabla f(x_k) - \nabla f(x^*)\right)^\top (x_k - x^*) \geq \tfrac{1}{L}\|\nabla f(x_k) - \nabla f(x^*)\|^2 \\
&\quad \text{and } \nabla f(x^*) = 0\text{)} \\
&\leq \|x_k - x^*\|^2 - \frac{1}{L^2}\|\nabla f(x_k)\|^2 \\
&\leq \|x_k - x^*\|^2.
\end{aligned}$$

Thus we get from (7), the following:

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\left(\frac{f(x_k) - f(x^*)}{\|x_1 - x^*\|}\right)^2. \tag{8}$$

Proof of Convergence III

In order to prove our result, we now call $r_k = f(x_k) - f(x^*)$ and $\gamma = \frac{1}{2L\|x_1 - x^*\|^2}$. Hence we have

$$r_{k+1} - r_k \leq -\gamma r_k^2.$$

Dividing by $r_k \cdot r_{k+1}$, we get

$$\frac{1}{r_k} - \frac{1}{r_{k+1}} \leq -\gamma \frac{r_k}{r_{k+1}},$$

that is

$$\frac{1}{r_{k+1}} \geq \frac{1}{r_k} + \gamma \frac{r_k}{r_{k+1}}.$$

Taking into account that $r_{k+1} \leq r_k$, we have

$$\frac{1}{r_{k+1}} \geq \frac{1}{r_k} + \gamma \frac{r_k}{r_{k+1}} \geq \frac{1}{r_k} + \gamma.$$

Summing up those inequalities, we get

$$\frac{1}{r_{k+1}} \geq \frac{1}{r_1} + \gamma k \geq \gamma k.$$

We can thus write

$$f(x_{k+1}) - f(x^*) \leq \frac{2L\|x_1 - x^*\|^2}{k}.$$

Comments

- This result says that gradient method has convergence rate $\mathcal{O}(1/k)$.
- We can calculate iterations needed to get gap lower or equal than ϵ.
- In practice, we want

$$f(x_{k+1}) - f(x^*) \leq \frac{c}{k} \leq \epsilon,$$

  with $c > 0$ depending on the values in previous Theorem
- We hence need a number of iterations of the order $\mathcal{O}(1/\epsilon)$.

Remark

- If we use exact line search we get the same rate.
- If we use Armijo line search to calculate α_k at each step, we obtain the same convergence rate with slightly different constant c (which depends on the parameters of the line search).
- Notice that $f(x_k) \to f(x^*)$ as $k \to \infty$.

Convergence analysis for the strongly convex case

Now, we consider the problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

where f is σ-strongly convex function with Lipschitz continuous gradient having constant L.

Remark

If f is σ-strongly convex function with Lipschitz continuous gradient having constant L, we have

$$\sigma \leq L.$$

We prove in the following theorem that gradient method with constant stepsize converges linearly.

Convergence result for the strongly convex case

Theorem (linear convergence for the strongly convex case)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a σ-strongly convex function with Lipschitz continuous gradient having Lipschitz constant $L > 0$. Gradient method with fixed stepsize $\alpha_k = 1/L$ satisfies:

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\sigma}{L}\right)^k (f(x_1) - f(x^*)).$$

Proof of Convergence (Stronlgy Convex Case)

Using Polyak-Lojasiewicz inequality

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \sigma(f(x) - f(x^*))$$

and the inequality

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2,$$

which still holds due to Lipschitz continuity of the gradient, we can write:

$$2\sigma\left(f(x_k) - f(x^*)\right) \leq \|\nabla f(x_k)\|^2 \leq 2L\left(f(x_k) - f(x_{k+1})\right).$$

Proof of Convergence (Stronlgy Convex Case)

If we call $r_k = f(x_k) - f(x^*)$, then we have

$$2\sigma r_k \leq 2L(r_k - r_{k+1}).$$

The last inequality can be rewritten as follows:

$$r_{k+1} \leq \left(1 - \frac{\sigma}{L}\right) r_k.$$

By induction, we have

$$r_{k+1} \leq \left(1 - \frac{\sigma}{L}\right)^k r_1,$$

which can be rewritten the following way

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\sigma}{L}\right)^k (f(x_1) - f(x^*)).$$

Comments

- Keep in mind that

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} \leq \sum_{i=0}^{\infty} x^i = \frac{1}{1-x},$$

  that is $1 - x \leq e^{-x}$, when $0 < x \leq 1$.

- Since $0 < \sigma/L \leq 1$, the previous result can be rewritten as follows

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\sigma}{L}\right)^k (f(x_1) - f(x^*)) \leq e^{-\frac{\sigma k}{L}} (f(x_1) - f(x^*)). \tag{9}$$

Comments II

- Gradient method has convergence rate $\mathcal{O}(c^k)$, with the constant $0 \le c < 1$ that depends on the values reported in previous Theorem.
- In the literature, the value L/σ is called the *condition number*.
- It is easy to see that, the higher is the condition number, the slower will be the convergence rate of our algorithm.

Comments III

- We can calculate the number of iterations needed to get an optimality gap lower or equal than ϵ.
- In practice, we want

$$f(x_{k+1}) - f(x^*) \leq c^k \leq \epsilon.$$

We hence have

$$k \log(c) \leq \log(\epsilon).$$

- Keeping in mind that $\tilde{c} = \dfrac{1}{\log(c)} < 0$ (if $c = e^{-\sigma/L}$, then $\tilde{c} = -L/\sigma$), we have

$$k \geq -\tilde{c} \log(\epsilon^{-1}).$$

- Thus we get a number of iterations of the order $\mathcal{O}(\log(1/\epsilon))$.

Improving the Rate

A slightly better rate can be obtained when using a different fixed stepsize.

Theorem [Linear Rate with Different Stepsize]

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a σ-strongly convex function with Lipschitz continuous gradient having Lipschitz constant $L > 0$. Gradient method with fixed stepsize $\alpha_k = 1/(\sigma + L)$ satisfies:

$$f(x_{k+1}) - f(x^*) \leq \frac{L}{2} \left(\frac{\frac{L}{\sigma} - 1}{\frac{L}{\sigma} + 1} \right)^{2k} \|x_1 - x^*\|^2.$$

We don't give here the proof. We just want to notice that in this case we have a rate order

$$\mathcal{O}\left(\left(1 - \frac{2}{\eta + 1} \right)^{2k} \right) \leq \mathcal{O}\left(e^{-\frac{4k}{\eta+1}} \right),$$

with $\eta = \frac{L}{\sigma}$, which is better than the rate obtained before.

Stopping Condition

- When minimizing strongly convex functions, we get a **Stopping Criterion** that guarantees to obtain an optimality gap under a given threshold.
- Using Polyak-Lojasiewicz inequality, we can write:

$$2\sigma\left(f(x_k) - f(x^*)\right) \leq \|\nabla f(x_k)\|^2.$$

If $f(x_k) - f(x^*) > \epsilon$ this implies

$$\sqrt{2\sigma\epsilon} < \|\nabla f(x_k)\|.$$

Thus we get that $\sqrt{2\sigma\epsilon} \geq \|\nabla f(x_k)\|$ implies

$$f(x_k) - f(x^*) \leq \epsilon.$$

Final Suggestions

- The gradient method is based on a simple idea and is very easy to implement.
- Each iteration is relatively cheap.
- Algorithm is very fast when dealing with well-conditioned and strongly convex problems.
- Calculation of σ and L not easy.
- Fixed stepsizes can hardly be used in practice!

Theory vs Practice

Theory

- Best choice: Fixed stepsize (e.g. $t = 1/L$)
- Elegant and simple way to prove results
- Similar worst case complexity for the 3 line search options

Practice

- Often hard to get L in practice
- Line search needs to be carefully chosen
- Other line searches might work better than fixed stepsize in the average case

Heavy Ball Method

- The *heavy ball* method is usually attributed to Polyak (1964).
- The iterates of gradient descent tend to bounce between the walls of narrow “valleys” on the objective surface.
- IDEA: add a momentum term to the gradient step, that is:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1}).$$

Heavy Ball Method: Example

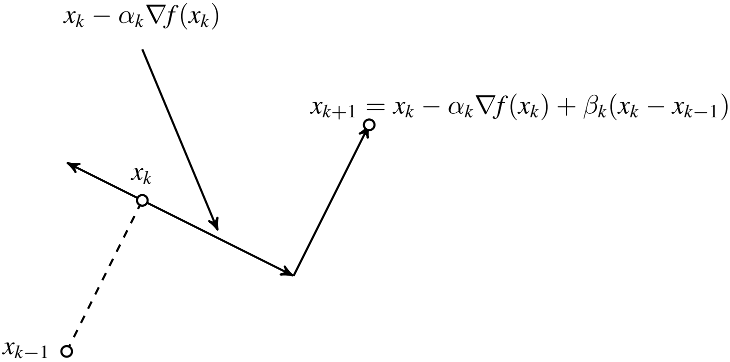

Figure: Illustration of the heavy ball method.

Heavy Ball Method: Details

- The term $x_k - x_{k-1}$, which is usually called *momentum*, nudges x_{k+1} in the direction of the previous step.
- Thanks to the momentum term, the method moves along the direction of the difference between the last two iterates (this is also called extrapolation step).
- It is possible to prove that heavy ball gets a better rate than the simple gradient (under same conditions seen before).

Heavy Ball Method: Comparison with Gradient Method

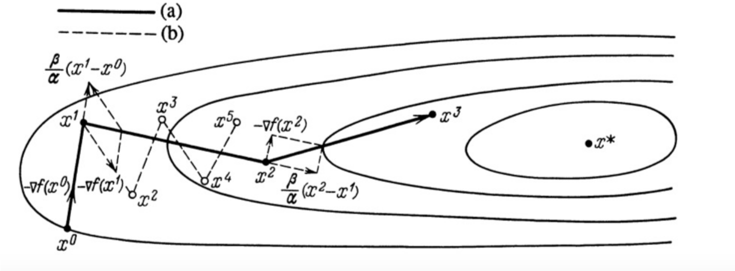

Figure: Comparison between heavy ball (a) and gradient method (b).

Conjugate Gradient Method

- In case we choose optimal parameters for α_k and β_k at each iteration, that is

$$(\alpha_k, \beta_k) \in \operatorname*{Argmin}_{\alpha \in \mathbb{R}\ \beta \in \mathbb{R}} f(x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})),$$

  the resulting method is an implementation of the so called *conjugate gradient* method.

- The method finds the minimum for a convex quadratic function in at most n iterations (where n is the dimension of our original problem).

Accelerated Gradient Method

- *Accelerated gradient method* proposed by Nesterov (1983).
- Structure of the algorithm very similar to heavy ball method.
- Generic iteration of the algorithm divided into two different steps:
  1. **Extrapolation step**: the method moves along the direction of the difference between the last two iterates, that is

     $$y_k = x_k + \beta_k(x_k - x_{k-1}),$$

     with β_k chosen depending on the properties of f;
  2. **Gradient step:** the method perform a gradient-like step at y_k to get x_{k+1}, that is

     $$x_{k+1} = y_k - \alpha_k \nabla f(y_k),$$

     with $\alpha_k = 1/L$ and L Lipschitz constant of the gradient.

Accelerated Gradient Method: Example

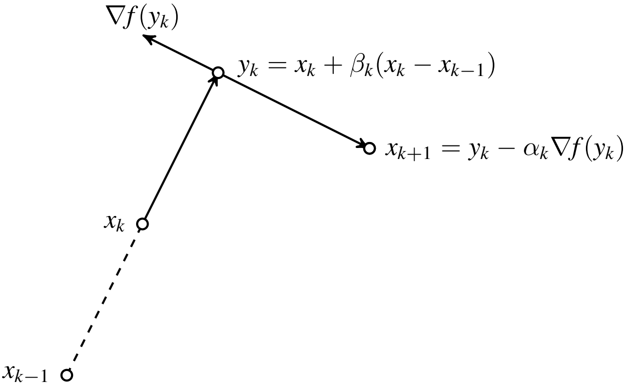

Figure: Illustration of the accelerated gradient method.

Comments

- Compared to the heavy ball, the accelerated gradient reverses the order of gradient calculation and extrapolation, and uses gradient calculated in y_k instead of gradient calculated in x_k.
- Method has better rates of convergence than the gradient.
- Possible to prove that accelerated gradient is the best we can get!

Comparison

	Gradient method	Nesterov's method
f convex, Lipschitz c. gradient	$\mathcal{O}(\frac{LD^2}{k})$	$\mathcal{O}(\frac{LD^2}{k^2})$
f σ-strongly convex, Lipschitz c. gradient	$\mathcal{O}\left(\left(\frac{\eta-1}{\eta+1}\right)^{2k}\right)$	$\mathcal{O}\left(\left(\frac{\sqrt{\eta}-1}{\sqrt{\eta}+1}\right)^{2k}\right)$

Table: Rates of convergence. Constants $\eta = \frac{L}{\sigma}$ and $D = \|x_1 - x^*\|$.