Chapter 4: Gradient Descent

$$\min_{x \in D} f(x)$$

- **first order method** : use only the gradient of the objective function $\nabla f(x)$
    - gradient descent method
- **second order method** : use the Hessian (the second order derivative), $\nabla^2 f(x)$, or its approximate, of the objective function.
    - Newton's method, quasi-Newton's method

Let's consider the unconstrained, smooth convex optimization

$$\min_{x} f(x)$$

We assume a few things about the function f :

- f is convex and differentiable (up to any order we need)
  -$\mathrm{dom}(f) = \mathbb{R}^n$, i.e., it has full domain

- We also assume here, like everywhere else in the course, that a solution exists (there are convex problems that get minimized out in infinity, but we assume we aren't in that case).
  Under this assumption, we denote the optimal criterion value by

$$f^* = \min_{x} f(x)$$

  with the solution at x^*.

> convexity assumption is mainly for some theoretical proof of many algorithms. Almost any algorithm can run without this assumption, although the performances can be drastically different for non-convex function f.

Gradient Descent (GD)

The **Gradient Descent** algorithm is then defined as follows:

1. Choose an initial point $x^{(0)} \in \mathbb{R}^n$
2. Repeat:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f\left(x^{(k-1)}\right), \quad \text{for } k = 1, 2, 3, \ldots$$

   where $t_k > 0$ is a sequence of pre-selected *step size*.
3. Stop at some point (i.e. *stopping criterion* – we talk this practicality later)

Above, after choosing some initial point $x^{(0)}$, we move it in the direction of the negative gradient (this points us in a direction where the function is decreasing) by some positive amount t_1, calling this x_1. And the same process is repeated.

Interpretation of GD

steepest descent direction

The *first-order Taylor expansion* of f gives us

$$f(x^{(k+1)}) \approx f(x^{(k)}) + \nabla f(x^{(k)})(x^{(k+1)} - x^{(k)}).$$

Write $x^{(k+1)} - x^{(k)} = tp_k, t > 0$ is the stepsize and p_k is the search direction. To enforce $f(x^{(k+1)}) < f(x^{(k)})$, we require

$$\nabla f(x^{(k)})(x^{(k+1)} - x^{(k)}) = t\nabla f(x^{(k)})p_k < 0$$

- **descent direction** p:

$$\nabla f(x) \cdot p < 0$$

- **steepest descent direction** p: $\nabla f(x) \cdot p \geq -\|\nabla f(x)\| \|p\|$ with the lowest value attained at

$$p = -\nabla f(x)$$

GD as a second-order Hessian approximationed by $\frac{I1}{\text{step size}}$

The second-order Taylor expansion of f at a given point x gives us

$$f(y) \approx f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)\nabla^2 f(x)(y-x)$$

Consider the *quadratic approximation* of f, replacing the Hessian matrix $\nabla^2 f(x)$ by a ***scalar matrix*** $\frac{1}{t}I$, we have

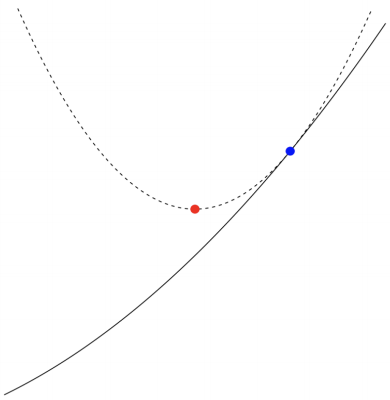
$$f(y) \approx f(x) + \nabla f(x)^T(y-x) + \frac{1}{2t}\|y-x\|_2^2$$

This is a convex quadratic, so we know we can minimize it just by setting its gradient in y to 0. Minimizing this w.r.t. y, we get

$$\begin{gathered}\frac{\partial f(y)}{\partial y} \approx \nabla f(x) + \frac{1}{t}(y-x) = 0 \\ \Longrightarrow y = x - t\nabla f(x)\end{gathered}$$

This gives us the above gradient descent update rule. In other words, gradients descent actually chooses the next point to minimize this approximated quadratic function .

Here the figure shows pictorially the interpretation. The dotted function shows the quadratic approximation with Hessian as a scalar matrix, and the red dot shows the minima of the quadratic approximation.



Blue point is x, red point is

$$x^+ = \underset{y}{\text{argmin}} \; f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2$$

GD: View point of *proximity operator*

We think of the GD

$$x^+ = \arg\min_y f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2$$

as the sum of two steps

- A linear approximation to $f$given by $f(x) + \nabla f(x)^T (y - x)$
- A *proximity* term to xx given by $\|y - x\|_2^2$ with weight $1/2t$.

Proximal Mapping

$$x \to \text{prox}_{h,t}(x) = \arg\min_z \frac{1}{2t} \|x - z\|_2^2 + h(z).$$

Here h is linear.

Topics of GD

- How to choose step sizes
- Continuous model - gradient flow

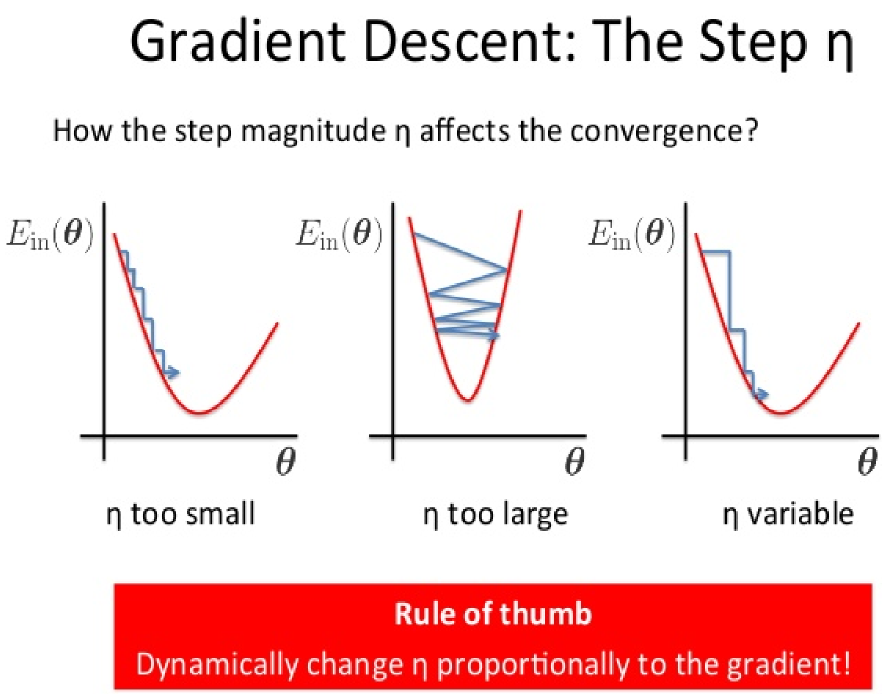
- Convergence analysis
- Practicality and stopping rule.

How to choose step sizes

In machine learning, the step size is also called *learning rate*, which is usually denoted by η.

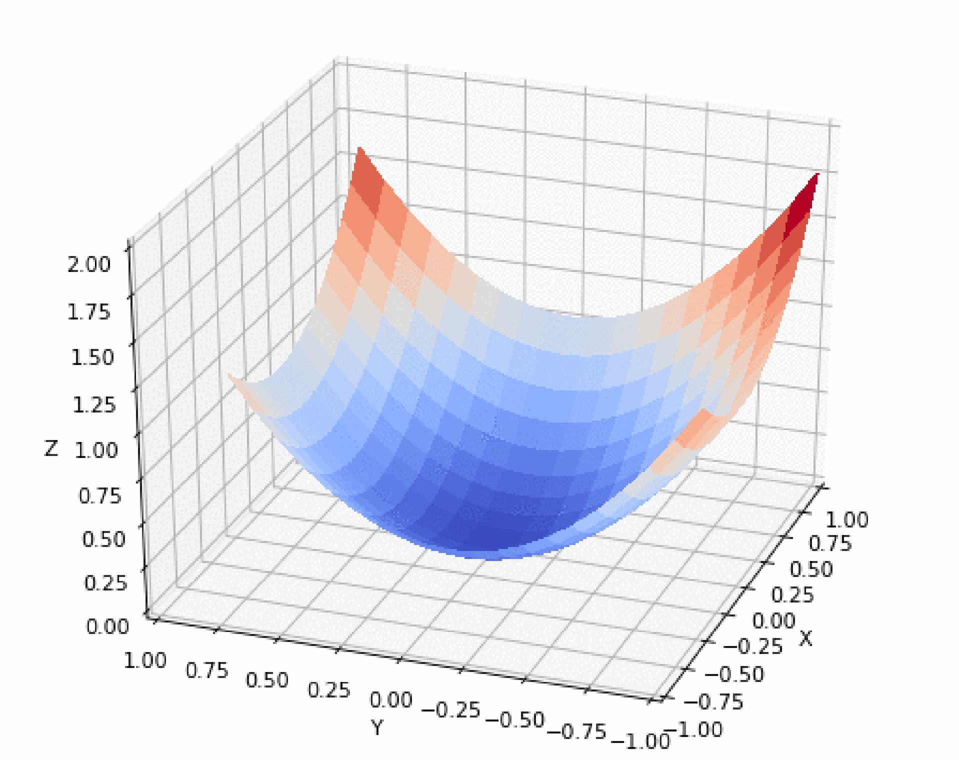
constant step size

- too small : convergence very slowly
- too large : move fast but may not converge – less stable



---

small and large step size :



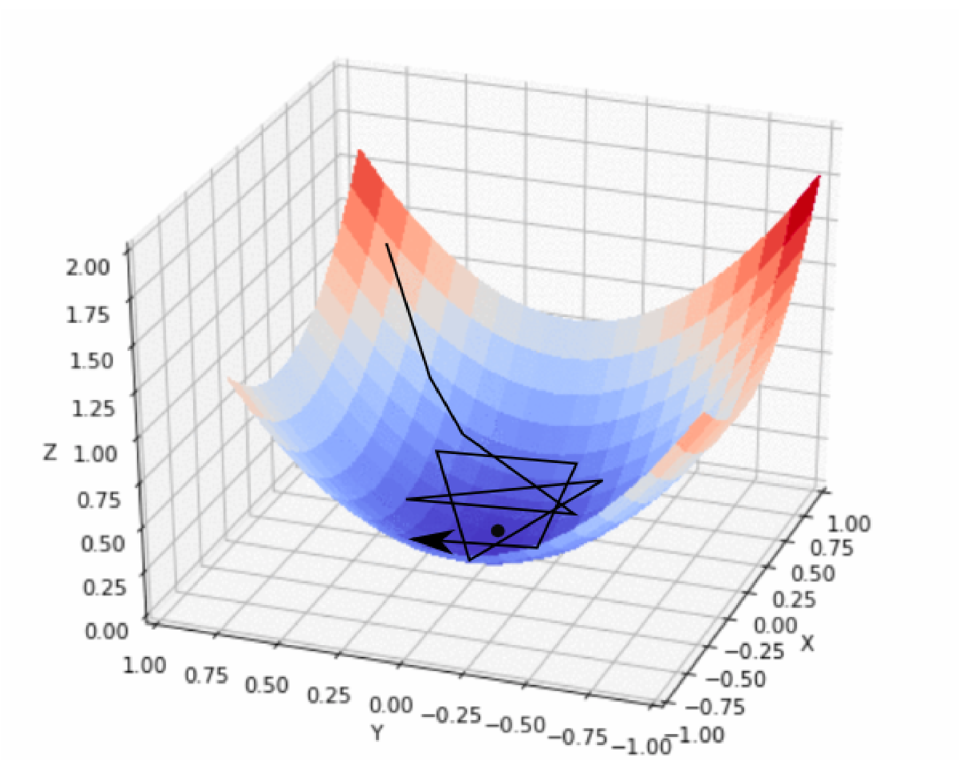


image credit: https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

Backtracking *line search*

One way to adaptively choose the step size is to use backtracking line search:

- First fix parameters $0 < \beta < 1$ and $0 < \alpha \leq 1/2$

- At each iteration, start with $t = t_{\text{init}} = 1$, and while

$$f(x - t\nabla f(x)) > f(x) - \alpha t \|\nabla f(x)\|_2^2$$

shrink

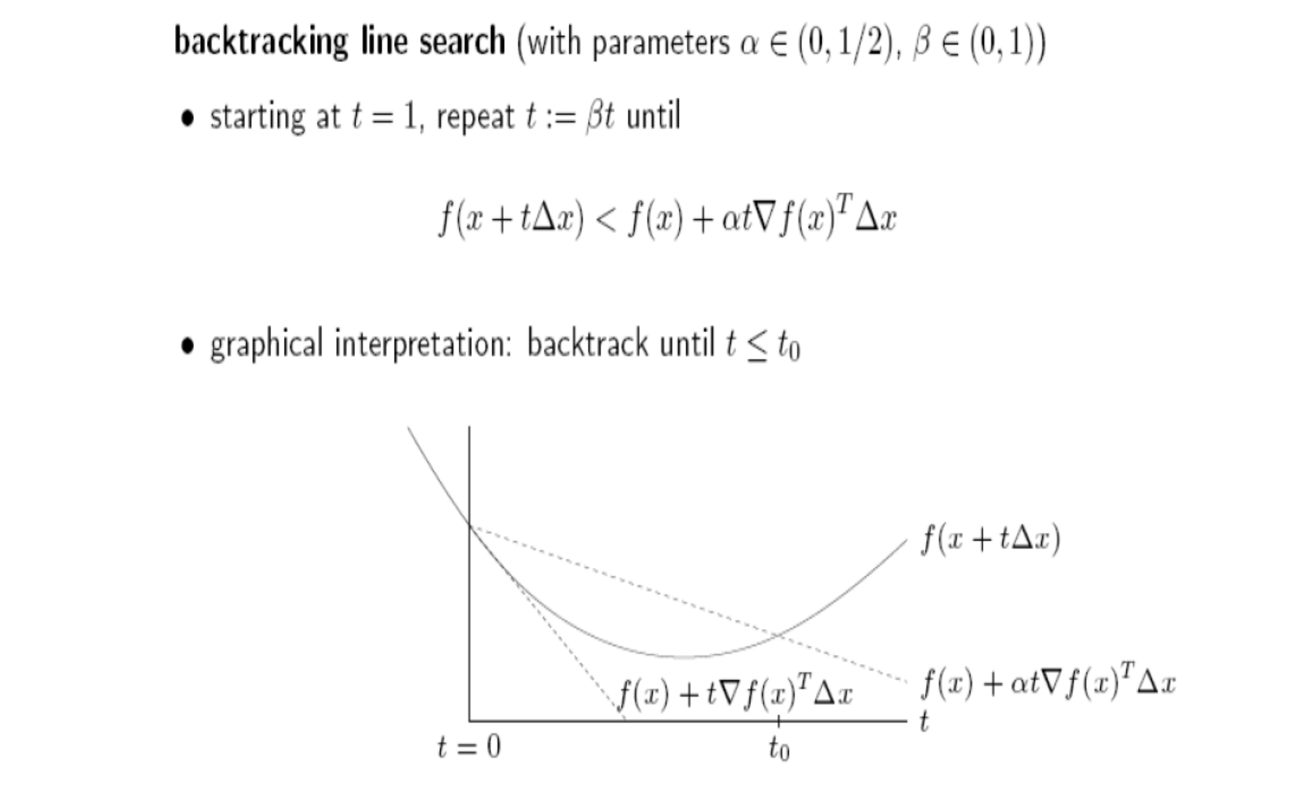
$$t \leftarrow \beta t.$$

> Recall if f is convex, then $f(y) \geq f(x) + \nabla f(x)^T(y - x)$. Let $y = x - t\nabla f(x)$, then
>
> $$f(x - t\nabla f(x)) > f(x) - t\|\nabla f(x)\|_2^2$$
>
> holds for any convex function.

- Else perform gradient descent update

$$x^+ = x - t\nabla f(x)$$

Simple and tends to work well in practice (further simplification: just take $\alpha = 1/2$)



For our case, $\Delta x = -\nabla f(x)$

The step size t_0 as the intersection point in the figure satisfies

$$f(x - t_0 \nabla f(x)) = f(x) - \alpha t_0 \|\nabla f(x)\|_2^2$$

Exact line search

We could also choose step to do the best we can along direction of negative gradient, called **exact line search**:

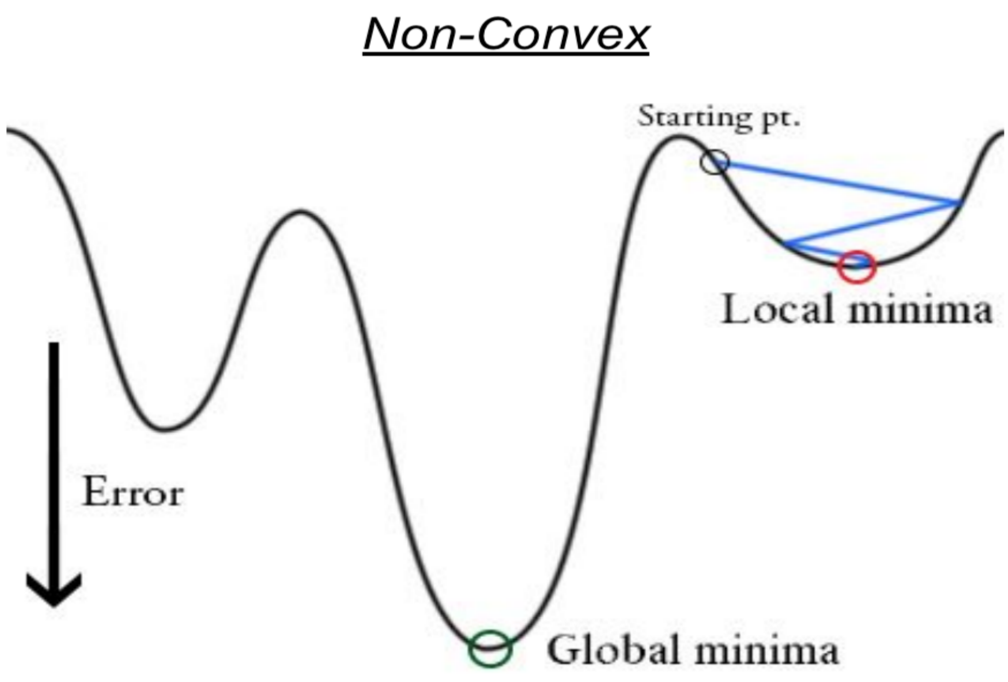
$$t = \arg\min_{s \geq 0} f(x - s\nabla f(x)).$$

- Usually not possible to do this minimization *exactly* (too expensive!).
- Approximations to exact line search are typically not as efficient as backtracking, and it's typically not worth it.

Pros and cons of gradient descent:

- Pro: simple idea, and each iteration is cheap (usually)
- Pro: fast for well-conditioned, strongly convex problems
- Con: can often be *slow*, because many interesting problems aren't strongly convex or well-conditioned.
- Con: can't handle *nondifferentiable* functions such as $\|x\|_1$. (We will discuss subgradient method later for this case)

GD for non-convex problem.



- differential initial points may go to different local minimum point
- other non-trivial slow-down phenomena (due to the shape of the function) in high dimension.

Continuous model of gradient descent

Rewrite the gradient descent starting with an initial x_0 as

$$x_{k+1} = x_k - \eta \nabla f(x) \Leftrightarrow \frac{x_{k+1} - x_k}{\eta} = -\nabla f(x)$$

where the step size is written as η and the variable t is used for other purpose.

As the step size η tends to zero, and think of the discrete sequence x_k as the approximate value of a continiuous function $X(t)$ as the continuous time $t_k = k\eta$. Note

$$\frac{x_{k+1} - x_k}{\eta} \approx \frac{X(t_{k+1}) - X(t_k)}{\eta} \to X'(t_k)$$

So, the continuous model of gradient descent is the **gradient flow** (an ordinary differential equation)

$$X'(t) = -\nabla f(X)$$

with $X(0) = x_0$

Property of gradient flow

If

$$X'(t) = -\nabla f(X),$$

then along the trajectory $X(t)$, the objective function does not increase:

$$\frac{d}{dt} f(X(t)) = \nabla f(X)^T X'(t) = -\|\nabla f(X(t))\|_2^2 \leq 0$$

with equality holds at the critical point $\nabla f(x) = 0$.

Example: $f(x) = \frac{1}{2}x^2$. Then GD with a constant step size η is

$$x_k - x_{k-1} = -\eta x_{k-1},$$

i.e.,

$$x_k = (1-\eta)x_{k-1} = (1-\eta)^k x_0.$$

Now fix a time interval $T > 0$ and for each integer K, let the step size $\eta = \eta^{(K)} = T/K$. Consider the discrete value x_K.

$$x_K = (1 - T/K)^K x_0 \to e^{-T} x_0 \triangleq X(T)$$

as $K \to +\infty$. Note $X(T)$ is the solution of the ODE $X'(t) = -X(t)$ with initial $X(0) = x_0$. So $x_K \to X(T)$ as the step size $\eta_k = T/K$ tends to zero.

In general, we have

$$\lim_{K\to\infty} \frac{1}{\eta^{(K)}} \max_{1\leq k\leq K} \|x_k - X(t_k)\| = c_T, \qquad t_k := k\eta^{(K)} = \frac{k}{K}T$$

where

- c_T is a constant depending on the time interval T.
- x_k is generated by the GD with the step size $\eta^{(K)} = T/K; t_k = k\eta = \frac{k}{K}T$ is the discrete time where x_k is defined.
- $X(t)$ is a continuously differentiable function solving the ODE

$$X'(t) = -\nabla f(X)$$

- $x_0 = X(t = 0)$

Convergence Theory of GD for convex function

Theorem 1 [convex with Lipschitz gradient]

If f is differentiale and convex with Lipschitz gradient:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for any } x, y$$

(Or when twice differentiable: $\nabla^2 f(x) \preceq LI$).

Then Gradient descent with fixed step size

$$\eta \leq \frac{1}{L}$$

(or backtracking with $\eta \leq \beta/L$) satisfies

$$f\left(x^{(k)}\right) - f^\star \leq \frac{\left\|x^{(0)} - x^\star\right\|_2^2}{2\eta k}$$

Theorem 2 [Lipschitz gradient+ strongly convex]

> Reminder: *(m-)stronlgy convex* means $f(x) - \frac{m}{2}\|x\|_2^2$ is convex for a positive m. For twice differentialbe function, it means $\nabla^2 f(x) \succeq mI$. m is the uniform bound of the smallest eigenvalues of the Hessian $\nabla^2 f(x)$ for all x.

If f is differentiale and convex and has L**-Lipschitz gradient, and** f **is** m**-strongly** positive, then the gradient descent with fixed step size

$$\eta \leq \frac{2}{m+L}$$

or with backtracking line search search satisfies

$$f\left(x^{(k)}\right) - f^\star \leq \gamma^k \frac{L}{2}\left\|x^{(0)} - x^\star\right\|_2^2$$

where $\gamma \in (0,1)$ is a constant dependent on m and L (roughly at the order $1 - \frac{m}{L}$).

> Rate under strong convexity is $O\left(\gamma^k\right)$, **exponentially fast** ! That is, it finds ϵ -suboptimal point in $O(\log(1/\epsilon))$ iterations:
>
> - $$\gamma^k \frac{L}{2}\left\|x^{(0)} - x^\star\right\|_2^2 \leq \epsilon \Rightarrow k \geq \frac{\log \frac{1}{\epsilon}}{-\log \gamma} + c_0, \quad \epsilon \ll 1$$
>
>   where c_0 is constant.
> - γ is roughly at the order $1 - \frac{m}{L}$; so $-\log\gamma \approx -\log(1 - \frac{m}{L}) \geq \frac{m}{L}$.
> - $k \geq \frac{\log \frac{1}{\epsilon}}{-\log \gamma} + c_0 \geq \frac{L}{m}\log\frac{1}{\epsilon} + c_0$, so the necessary nubmer of steps is
>
> $$\frac{L}{m}\log\frac{1}{\epsilon}$$
>
>   for $f(x_k) - f^\star \leq \epsilon$.

linear convergence

$$f\left(x^{(k)}\right) - f^\star \leq \gamma^k \frac{L}{2}\left\|x^{(0)} - x^\star\right\|_2^2$$

implies

$$\log(f\left(x^{(k)}\right) - f^\star) \leq k\log\gamma + \log\left(\frac{L}{2}\left\|x^{(0)} - x^\star\right\|_2^2\right)$$

So the semi-log plot

$$\log(f\left(x^{(k)}\right) - f^\star) \text{ v.s. } k$$

is linear with the (negative) slope $\log \gamma$.

Definition of Condition number

The conditions in Theorem 2 for twice differentiable function is summarized as

$$mI \preceq \nabla^2 f(x) \preceq LI$$

- $\gamma \approx 1 - \frac{m}{L}$: the smaller γ, the faster convergence
- m: the smallest eigenvalue
- L: the largest eigenvalue
- **condition number** is then defined by

$$\text{cond} \triangleq \frac{L}{m} = \frac{\max \text{eigenvalue}}{\min \text{eigenvalue}}$$

- The larger the condition number, the larger γ, the slower the convergence, meaning the problem is harder for GD to solve: Theorem 2 indiecates that

$$k \geq \text{cond} \times \log \frac{1}{\epsilon}$$

  in order to have $f(x_k) - f^\star \leq \epsilon$.

Practicality

Stopping rule

stop when $\|\nabla f(x)\|_2$ is small. Recall

$$\nabla f\left(x^\star\right) = 0$$

at solution $x^\star$

- This is not sufficient for convergence to the optimal minimum point in theory, unless f is **strongly convex**.
- It is important to "visualize" your training process by plotting the curve $f(x_k)$ v.s. k.
- try different values of constant step size in practice

Justification

Important inequalities for strongly convex function

Recall that a differential function f is **strongly convex** with constant m is equivalent to any of the following statements

1. $(\nabla f(x) - \nabla f(y))^T(x-y) \geq m\|x-y\|_2^2$ for all x, y
2. $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2$ for all x, y

We shall show that the gradient can bounds the error in the optimal point and the optimal value: for any $x \in \mathrm{dom}(f)$,

- $$\|\nabla f(x)\| \geq m\|x-x^\star\|$$
- $$\|\nabla f(x)\|^2 \geq 2m(f(x)-f^\star)$$

Proof:
For the first equivalent condition:
4. Let $y = x^\star$, then (since $\nabla f(x^\star) = 0$),

$$\nabla f(x)^T(x-x^\star) = (\nabla f(x) - \nabla f(x^\star))^T(x-x^\star) \geq m\|x-x^\star\|_2^2$$

$$\implies \qquad \|\nabla f(x)\| \geq m\|x-x^\star\|$$

> *This inequality shows that if the gradient is small at a point, then the point is nearly optimal.*

For the second condition: The RHS has a minimal value

$$\min_y \left[f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2\right]$$

as $y = \tilde{y} = x - \frac{1}{m}\nabla f(x)$. So

$$
\begin{aligned}
& f(x)+\nabla f(x)^T(y-x)+\frac{m}{2}\|y-x\|_2^2 \\
\geq & f(x)+\nabla f(x)^T(\tilde{y}-x)+\frac{m}{2}\|\tilde{y}-x\|_2^2 \\
= & f(x)-\frac{1}{2m}\|\nabla f(x)\|^2
\end{aligned}
$$

Then

$$
\Longrightarrow \qquad f(y) \geq f(x)-\frac{1}{2m}\|\nabla f(x)\|^2, \quad \forall x, y
$$

$$
\Longrightarrow \qquad f(x^\star) \geq f(x)-\frac{1}{2m}\|\nabla f(x)\|^2
$$

$$
f(x)-f^\star \leq \frac{1}{2m}\|\nabla f(x)\|^2
$$

> *This inequality shows that if the gradient is small at a point, then the value of f is nearly optimal.*

If we have

$$
\|\nabla f(x_k)\|_2 \leq \sqrt{2m\epsilon}
$$

as the **stopping rule**, then we have

- $$
f(x_k)-f^\star \leq \frac{1}{2m}\|\nabla f(x_k)\|_2^2=\epsilon
$$

  This justifies our use of the stopping rule for small $\|\nabla f(x_k)\|_2$ for the strongly convex function.

homework

Let $f(x)=\frac{L}{2}x^2$, $x \in \mathbb{R}$. Show that the bound in Theorem 2 for the GD with a fixed step size η is sharp, i.e., there is γ such that the equality holds

$$f\left(x^{(k)}\right) - f^\star = \gamma^k \frac{L}{2} \left\| x^{(0)} - x^\star \right\|_2^2$$

Find the value of γ.

> answer : $\gamma = 1 - \eta L.$

References and further reading

- S. Boyd and L. Vandenberghe (2004), “Convex optimization”, Chapter 9
- T. Hastie, R. Tibshirani and J. Friedman (2009), “The elements of statistical learning”, Chapters 10 and 16
- Y. Nesterov (1998), “Introductory lectures on convex optimization: a basic course”, Chapter 2
- L. Vandenberghe, Lecture notes for EE 236C, UCLA, Spring 2011-2012