5. Convex optimization

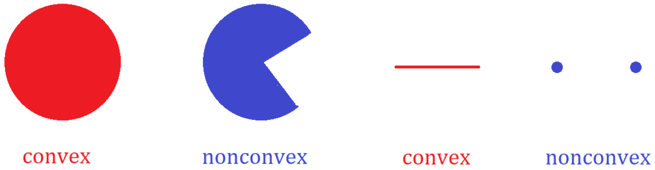
- Convex sets, functions
- Gradient descent
- Gradient based methods

Convex sets, functions

Convex set

$\mathcal{S} \subset \mathbb{R}^d$ is a convex set if

$$\forall \theta, \nu \in \mathcal{S}, \quad \alpha\theta + (1-\alpha)\nu \in \mathcal{S}, \quad 0 \leq \alpha \leq 1$$



Example: linear subspace

$$\mathcal{S} \text{ is convex if } \forall \theta, \nu \in \mathcal{S}, \quad \alpha\theta + (1-\alpha)\nu \in \mathcal{S}, \quad 0 \le \alpha \le 1$$

- $\mathcal{S}$ is a <u>linear subspace</u> if

$$\forall \theta, \nu \in \mathcal{S},\ \forall \alpha, \beta \in \mathbb{R}, \quad \alpha\theta + \beta\nu \in \mathcal{S}$$

- Convex from definition
- Example: null space is convex. Proof:

$$A\theta = 0,\ A\theta = 0, \Rightarrow A(\alpha\theta + (1 - \alpha\theta)) = 0$$

- Example: solution space is convex. Proof:

$$X\theta = y,\ X\nu = y, \Rightarrow X(\alpha\theta + (1-\alpha)\nu) = \alpha y + (1-\alpha)y = y$$

Example: half-space

$$\mathcal{S} \text{ is convex if } \forall \theta, \nu \in \mathcal{S}, \quad \alpha\theta + (1-\alpha)\nu \in \mathcal{S}, \quad 0 \le \alpha \le 1$$

- $\mathcal{S}$ is a linear halfspace if it can be written as

$$\mathcal{S} = \{\theta : X\theta \ge y\}$$

  for some X, y

- Convex?

- Example: nonnegative orthant $\mathcal{S} = \{\theta : \theta_i \ge 0, \ \forall i\}$

$$\theta_i \ge 0, \quad \nu_i \ge 0 \Rightarrow \alpha\theta_i + (1-\alpha)\nu_i \ge 0$$

[redacted]

Example: half-space

$$\mathcal{S} \text{ is convex if } \forall \theta, \nu \in \mathcal{S}, \quad \alpha\theta + (1-\alpha)\nu \in \mathcal{S}, \quad 0 \leq \alpha \leq 1$$

- $\mathcal{S}$ is a linear halfspace if it can be written as

$$\mathcal{S} = \{\theta : X\theta \geq y\}$$

  for some X, y

- Convex? Yes! Proof: plug into definition of convexity.
- Example: nonnegative orthant $\mathcal{S} = \{\theta : \theta_i \geq 0, \ \forall i\}$

$$\theta_i \geq 0, \quad \nu_i \geq 0 \Rightarrow \alpha\theta_i + (1-\alpha)\nu_i \geq 0$$

  since each term is nonnegative.

Example: Set of positive (semi)definite matrices

$\mathcal{S}$ is convex if $\forall \theta, \nu \in \mathcal{S}, \quad \alpha\theta + (1-\alpha)\nu \in \mathcal{S}, \quad 0 \leq \alpha \leq 1$

- The set of <u>positive semidefinite (PSD) matrices</u> is defined as

$$\mathcal{S} = \{X : u^T X u \geq 0,\ \forall u\}$$

- The set of <u>positive definite (PD) matrices</u> is defined as

$$\mathcal{S} = \{X : u^T X u > 0,\ \forall u \neq 0\}$$

- Convex?

Example: Set of positive (semi)definite matrices

$\mathcal{S}$ is convex if $\forall \theta, \nu \in \mathcal{S}, \quad \alpha\theta + (1-\alpha)\nu \in \mathcal{S}, \quad 0 \leq \alpha \leq 1$

- The set of positive semidefinite (PSD) matrices is defined as

$$\mathcal{S} = \{X : u^T X u \geq 0,\ \forall u\}$$

- The set of positive definite (PD) matrices is defined as

$$\mathcal{S} = \{X : u^T X u > 0,\ \forall u \neq 0\}$$

- Convex? Yes! if $u^T X u \geq 0$ and $u^T Y u \geq 0$ then

$$u^T(\alpha X + (1-\alpha)Y)u = \alpha u^T X u + (1-\alpha)u^T Y u \geq 0$$

Proof for PD matrices is basically the same

Other important properties

- The intersection of convex sets is convex

$$\mathcal{S}_1, \mathcal{S}_2 \text{ convex } \Rightarrow \mathcal{S}_1 \cap \mathcal{S}_2 \text{ convex}$$

- The union of sets is usually not convex

Example: $\{0\}$, $\{1\}$, ..., $\{n\}$ each convex, set of integers $\{1, 2, ..., n\}$ is not convex

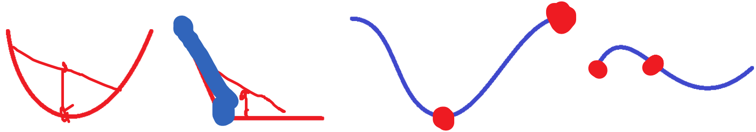
- Affine transformations preserve convexity

$$\mathcal{S} \text{ convex } \Rightarrow A\mathcal{S} + b = \{Ax + b : x \in \mathcal{S}\} \text{ convex}$$

Convex function

A function $f : \mathcal{Z} \to \mathbb{R}$ is convex if

$$\forall \theta, \nu \in \mathcal{Z}, \quad f(\alpha\theta + (1-\alpha)\nu) \leq \alpha f(\theta) + (1-\alpha) f(\nu) \; \forall 0 \leq \alpha \leq 1$$



convex convex nonconvex nonconvex

Operations that preserve convexity

- **Affine transformation**

$$f(x) \text{ is convex } \Rightarrow g(w) = f(Aw + b) \text{ is convex.}$$

- **Pointwise max and supremum**

$$f(x) = \max_i f_i(x), \qquad g(x) = \sup_{s \in \mathcal{S}} g_s(x)$$

are convex if each f_i, each g_s are convex ($\mathcal{S}$ may not be convex)

- **Minimization**

$$g(x) = \inf_{y \in \mathcal{C}} f(x, y)$$

is convex if f is convex in (x, y), $\mathcal{C}$ are convex

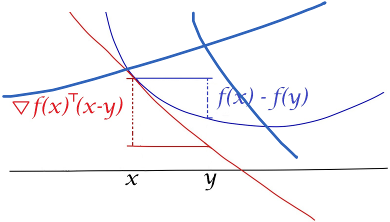
First order condition

- Recall the gradient operator $\nabla f : \mathcal{Z} \to \mathbb{R}^d$

$$\nabla f(\theta) = \begin{bmatrix} \partial f / \partial \theta_1 \\ \vdots \\ \partial f / \partial \theta_d \end{bmatrix}$$

- If $f : \mathcal{Z} \to \mathbb{R}$ is 1-differentiable, then f is convex if and only if

$$f(\theta) - f(\nu) \geq \nabla f(\nu)^T (\theta - \nu)$$



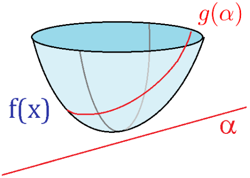
Second order condition

- Recall the Hessian operator $\nabla f : \mathcal{Z} \to \mathbb{R}^{d\times d}$

$$\nabla^2 f(\theta) = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1^2} & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_d^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \theta_d \partial \theta_1} & \frac{\partial^2 f}{\partial \theta_d \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_d^2} \end{bmatrix}$$

- If $f : \mathcal{Z} \to \mathbb{R}$ is 2-differentiable, then $f : \mathcal{Z} \to \mathbb{R}$ is convex if and only if

$$\nabla^2 f(\theta) \text{ is PSD for all } \theta \in \mathcal{Z}$$



$g''(\alpha) \geq 0$ for all α

Example: Regularized linear regression

$$f(\theta) = \frac{1}{2}\|X\theta - y\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2$$

- Twice differentiable $\rightarrow$ compute Hessian, check if convex

$$\nabla f(\theta) = X^T(X\theta - y) + \lambda\theta, \qquad \nabla^2 f(\theta) = X^T X + \lambda I$$

- Hessian does not depend on θ!
- Positive (semi)definite?

Example: Regularized linear regression

$$f(\theta) = \frac{1}{2}\|X\theta - y\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2$$

- Twice differentiable $\rightarrow$ compute Hessian, check if convex

$$\nabla f(\theta) = X^T(X\theta - y) + \lambda\theta, \qquad \nabla^2 f(\theta) = X^T X + \lambda I$$

- Hessian does not depend on θ!
- Positive (semi)definite? Ans: yes, PD if $\lambda > 0$, at least PSD if $\lambda = 0$

$$u^T(X^T X + \lambda I)u = \|Xu\|_2^2 + \lambda\|u\|_2^2 \geq 0$$

Global vs local optimality, stationary points

Consider the unconstrained minimization problem for f everywhere differentiable

$$\underset{\theta}{\text{minimize}}\ f(\theta)$$

- Any θ where $\nabla f(\theta) = 0$ is a stationary point
- If f is convex, then $\nabla f(\theta) = 0$ implies θ is a global minimum

$$\forall \theta', \quad f(\theta) \leq f(\theta')$$

- θ is a local minimum if

$$\forall \theta' : \|\theta - \theta'\| \leq \epsilon, \quad f(\theta) \leq f(\theta')$$

- If θ is stationary and $\nabla^2 f(\theta)$ is PD, θ is a local minimum

Quasiconvex functions

- A sublevel set of $f(x)$ is defined as

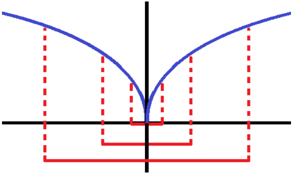
$$\mathcal{S}_\alpha = \{x : f(x) \leq \alpha\}.$$

  Every convex function has only convex sublevel sets.

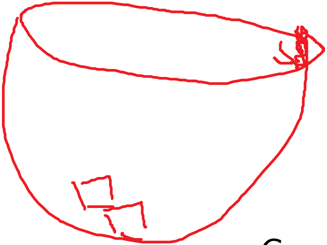
- If a function has convex sublevel sets but is not convex, it is quasiconvex.

Quasiconvex functions

Example: $f(x) = \sqrt{|x|}$



- Function is not convex
- Sublevel sets are closed intervals $\Rightarrow$ convex.
- What happens when you use a gradient method? Step size choice?



Gradient descent

Gradient descent

$$\underset{\theta}{\text{minimize}} \quad f(\theta) := \sum_{i=1}^{m} f_i(\theta), \qquad f_i \text{ is differentiable everywhere}$$

- Gradient descent step: given stepsize $\alpha > 0$,

$$\theta^{(k+1)} = \theta^{(k)} - \alpha \underbrace{\sum_{i=1}^{m} \nabla f_i(\theta^{(k)})}_{=\nabla f(\theta^{(k)})}$$

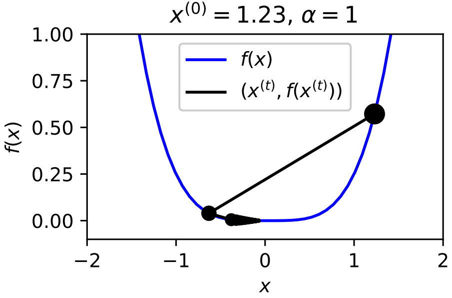
- If m is very large, each step pretty slow
- Difference between sampling $m = 100$ vs $m = 1$ billion?

see whiteboard

Quartic function

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad \frac{x^4}{4}$$

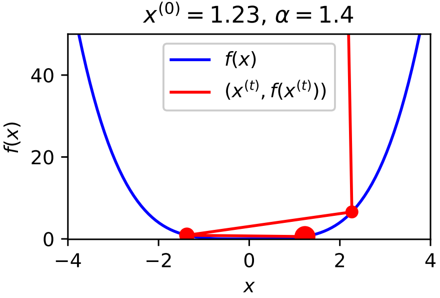
Gradient descent: $x^{(k+1)} = x^{(k)} - \alpha(x^{(k)})^3$



Quartic function

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad \frac{x^4}{4}$$

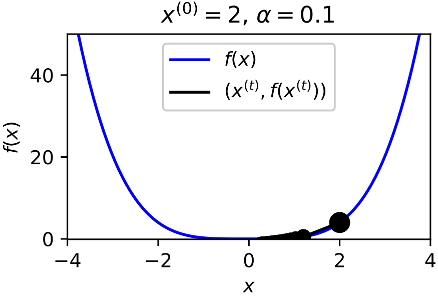
Gradient descent: $x^{(k+1)} = x^{(k)} - \alpha (x^{(k)})^3$



Quartic function

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad \frac{x^4}{4}$$

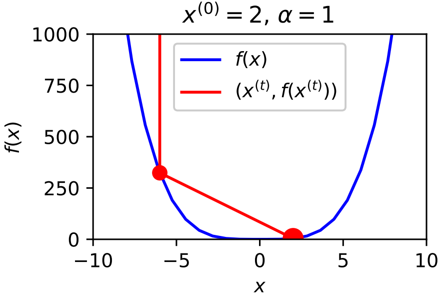
Gradient descent: $x^{(k+1)} = x^{(k)} - \alpha(x^{(k)})^3$



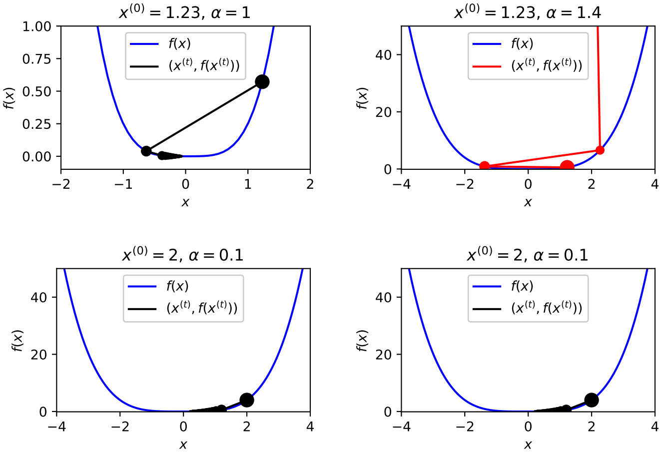
Quartic function

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad \frac{x^4}{4}$$

Gradient descent: $x^{(k+1)} = x^{(k)} - \alpha (x^{(k)})^3$



Quartic function

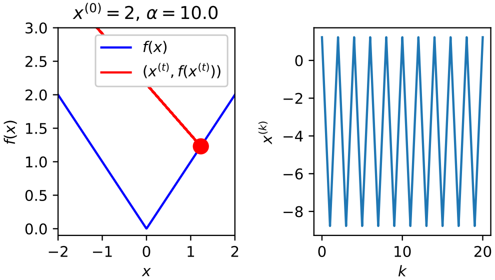


Somehow, convergence vs divergence depends on step size <u>and</u> initial value!

Absolute value function

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad |x|$$

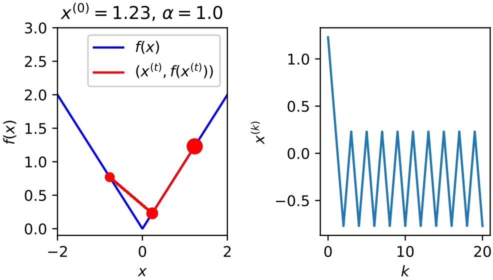
Gradient descent: $x^{(k+1)} = x^{(k)} - \alpha\, \mathbf{sign}(x^{(k)})$



Absolute value function

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad |x|$$

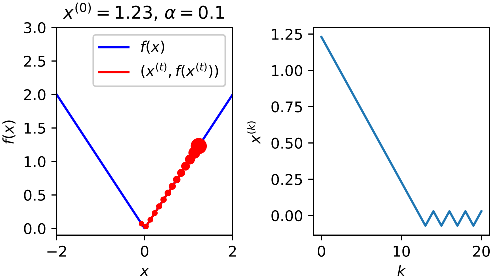
Gradient descent: $x^{(k+1)} = x^{(k)} - \alpha\,\mathbf{sign}(x^{(k)})$



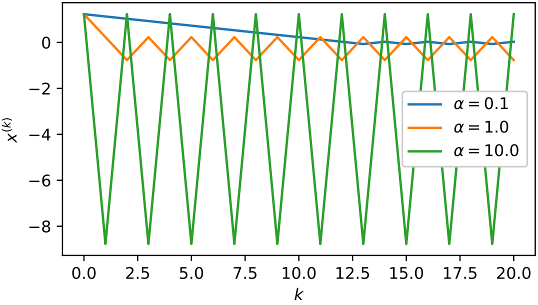
Absolute value function

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad |x|$$

Gradient descent: $x^{(k+1)} = x^{(k)} - \alpha\, \mathbf{sign}(x^{(k)})$



Absolute value function



Doesn't converge no matter what we do!

Smoothness

We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if its gradient is L-Lipschitz

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

We say that $f(x)$ is smooth if some L exists.

- Is $f(x) = x^4/4$ smooth?
- Is $f(x) = |x|$ L-smooth?
- Is $f(x) = x^2/2$ L-smooth?

Smoothness

We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is <u>L-smooth</u> if its gradient is L-Lipschitz

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

We say that $f(x)$ is <u>smooth</u> if some L exists.

- Is $f(x) = x^4/4$ smooth? Ans: No, trouble at $|x| \to +\infty$
- Is $f(x) = |x|$ L-smooth?
- Is $f(x) = x^2/2$ L-smooth?

Smoothness

We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is <u>L-smooth</u> if its gradient is L-Lipschitz

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

We say that $f(x)$ is <u>smooth</u> if some L exists.

- Is $f(x) = x^4/4$ smooth? Ans: No, trouble at $|x| \to +\infty$
- Is $f(x) = |x|$ L-smooth? Ans: No, trouble at $x = 0$
- Is $f(x) = x^2/2$ L-smooth?

Smoothness

We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if its gradient is L-Lipschitz

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

We say that $f(x)$ is smooth if some L exists.

- Is $f(x) = x^4/4$ smooth? Ans: No, trouble at $|x| \to +\infty$
- Is $f(x) = |x|$ L-smooth? Ans: No, trouble at $x = 0$
- Is $f(x) = x^2/2$ L-smooth? Ans: Yes! $L = 1$

Descent lemma

- If f is L-smooth, then picking $\alpha < 2/L$ guarantees descent:

$$f(x - \alpha \nabla f(x)) \leq f(x).$$

- Furthermore, the gradient goes to 0

$$x^{(t+1)} = x^{(t)} - \alpha \nabla f(x^{(t)}), \qquad \cancel{\|x^{(t)}\|_2} \overset{t\to\infty}{\to} 0$$

$$\|\nabla f(x^{(t)})\|_2 \to 0$$

- In particular, we do not require f to be convex.

Proof of descent lemma

Step one: Alternative form of Lipschitz smoothness

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

implies

$$f(x) \leq f(y) + \nabla f(y)^T(x - y) + \frac{L}{2}\|x - y\|_2^2$$

Proof in 1-D

$$\begin{aligned} f(y) - f(x) &= \int_x^y f'(u)du \\ &= \underbrace{\int_x^y f'(y)du}_{=f'(y)(y-x)} + \int_x^y \underbrace{f'(u) - f'(y)}_{\leq |f'(u) - f'(y)|} du \end{aligned}$$

Since

$$|f'(u) - f'(y)| \leq |f'(x) - f'(y)| \leq \frac{L}{2}|x - y|$$

result follows

Proof of descent lemma

Step two: Plug in alternate Lipschitz smoothness

$$\begin{aligned} f(x - \alpha \nabla f(x)) &\leq f(x) + \nabla f(x)^T(-\alpha \nabla f(x)) + \frac{L}{2}\|\alpha \nabla f(x)\|_2^2 \\ &= \alpha\left(\frac{L\alpha}{2} - 1\right)\|\nabla f(x)\|_2^2 + f(x) \end{aligned}$$

- As long as ???, each step is a guaranteed descent step
- The best choice of step size, based on this bound, is ???

Proof of descent lemma

Step two: Plug in alternate Lipschitz smoothness

$$\begin{aligned} f(x - \alpha \nabla f(x)) &\leq f(x) + \nabla f(x)^T(-\alpha \nabla f(x)) + \frac{L}{2}\|\alpha \nabla f(x)\|_2^2 \\ &= \alpha \left(\frac{L\alpha}{2} - 1 \right) \|\nabla f(x)\|_2^2 \end{aligned}$$

- As long as $\alpha < 2/L$, each step is a guaranteed descent step
- The best choice of step size, based on this bound, is ???

Proof of descent lemma

Step two: Plug in alternate Lipschitz smoothness

$$
\begin{aligned}
f(x-\alpha\nabla f(x)) &\leq f(x)+\nabla f(x)^T(-\alpha\nabla f(x))+\frac{L}{2}\|\alpha\nabla f(x)\|_2^2 \\
&= \underbrace{\alpha\left(\frac{L\alpha}{2}-1\right)}\|\nabla f(x)\|_2^2 + f(x)
\end{aligned}
$$

- As long as $\alpha < 2/L$, each step is a guaranteed descent step
- The best choice of step size, based on this bound, is $\alpha = 1/L$

Corollary of descent lemma

Gradient norm to 0:

$$\sum_{t=0}^{T} f(x^{(t+1)}) - f(x^{(t)}) \leq \alpha \left(\frac{L\alpha}{2} - 1 \right) \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|_2^2$$

Corollary of descent lemma

Gradient norm to 0:

$$\sum_{t=0}^{T} f(x^{(t+1)}) - f(x^{(t)}) \leq \alpha \left(\frac{L\alpha}{2} - 1 \right) \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|_2^2$$

$\Downarrow$ (Telescoping)

Corollary of descent lemma

Gradient norm to 0:

$$\sum_{t=0}^{T} f(x^{(t+1)}) - f(x^{(t)}) \leq \alpha \left(\frac{L\alpha}{2} - 1 \right) \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|_2^2$$

$\Downarrow$ (Telescoping)

$$f(x^{(0)}) - f(x^{(T)}) \geq \alpha \left(\frac{L\alpha}{2} - 1 \right) \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|_2^2$$

Corollary of descent lemma

Gradient norm to 0:

$$\sum_{t=0}^{T} f(x^{(t+1)}) - f(x^{(t)}) \leq \alpha \left(\frac{L\alpha}{2} - 1 \right) \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|_2^2$$

$\Downarrow$ (Telescoping)

$$f(x^{(0)}) - f(x^*) \geq f(x^{(0)}) - f(x^{(T)}) \geq \alpha \left(\frac{L\alpha}{2} - 1 \right) \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|_2^2$$

where $x^* =$ limit point of $x^{(t)}$

Corollary of descent lemma

Gradient norm to 0:

$$\sum_{t=0}^{T} f(x^{(t+1)}) - f(x^{(t)}) \leq \alpha \left(\frac{L\alpha}{2} - 1 \right) \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|_2^2$$

$\Downarrow$ (Telescoping)

$$\underbrace{f(x^{(0)}) - f(x^*)}_{\text{Bounded}} \geq f(x^{(0)}) - f(x^{(T)}) \geq \alpha \left(\frac{L\alpha}{2} - 1 \right) \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|_2^2$$

where $x^* =$ limit point of $x^{(t)}$

Corollary of descent lemma

Gradient norm to 0:

$$\sum_{t=0}^{T} f(x^{(t+1)}) - f(x^{(t)}) \leq \alpha \left(\frac{L\alpha}{2} - 1 \right) \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|_2^2$$

$\Downarrow$ (Telescoping)

$$\underbrace{f(x^{(0)}) - f(x^*)}_{\text{Bounded}} \geq f(x^{(0)}) - f(x^{(T)}) \geq \alpha \left(\frac{L\alpha}{2} - 1 \right) \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|_2^2$$

where $x^* =$ limit point of $x^{(t)}$

$$\Rightarrow \|\nabla f(x^{(t)})\|_2 \to 0$$

More details

- Without further details, gradient descent on an L-smooth problem reaches a stationary point at iteration complexity $O(1/\epsilon)$

  (Read as "it takes $O(1/\epsilon)$ iterations for $f(x^{(t)}) - f(x^*) = \epsilon$")

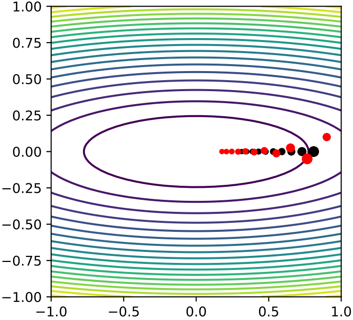
- With acceleration, the iteration complexity reduces to $O(1/\sqrt{\epsilon})$. This is the best you can do using only 1st-order information [1]

- If problem is additionally strongly convex, rates may be much better

[1] See Nesterov '83 or book Introductory Lectures on Convex Optimization

Ellipse

$$\underset{x}{\text{minimize}} \quad x_1^2 + 10000 \cdot x_2^2$$



(Go to demo)

Strong convexity

- $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if   for mu > 0

$$f(x) - f(y) \geq \nabla f(y)^T (x - y) + \frac{\mu}{2}\|x - y\|_2^2$$

- Compare with L-smooth:

$$f(x) - f(y) \leq \nabla f(y)^T (x - y) + \frac{L}{2}\|x - y\|_2^2$$

- Hessian bound: f is μ-strongly convex and L-smooth if for all x,

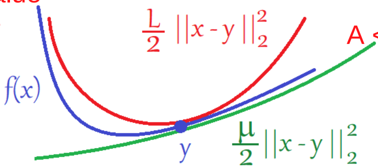
$$\mu I \preceq \nabla^2 f(x) \preceq LI$$

squiggly <= means

A <= B iff (B-A) is positive semidefinite

not elementwise

mu is smallest eigenvalue
L is largest eigenvalue



Convergence under strong convexity

If f is L-smooth and μ-strongly convex

- There is a unique stationary point x^*

Convergence under strong convexity

If f is L-smooth and μ-strongly convex

- There is a unique stationary point x^*

  Proof Suppose that $\nabla f(x) = \nabla f(y) = 0$

  Then

$$\begin{aligned} f(x) - f(y) &\geq \frac{\mu}{2}\|x - y\|_2^2 \\ f(y) - f(x) &\geq \frac{\mu}{2}\|x - y\|_2^2 \\ \Rightarrow 0 &\geq \mu\|x - y\|_2^2 \end{aligned}$$

Convergence under strong convexity

If f is L-smooth and μ-strongly convex

- There is a unique stationary point x^*
- The constant $\kappa = L/\mu$ is often called the condition number of f (Compare with condition number of a matrix)

Convergence under strong convexity

If f is L-smooth and μ-strongly convex

- There is a unique stationary point x^*
- The constant $\kappa = L/\mu$ is often called the condition number of f
  (Compare with condition number of a matrix)
- Gradient descent converges at rate $O(\kappa \log(\epsilon))$

Convergence under strong convexity

If f is L-smooth and μ-strongly convex

- There is a unique stationary point x^*
- The constant $\kappa = L/\mu$ is often called the condition number of f
  (Compare with condition number of a matrix)
- Gradient descent converges at rate $O(\kappa \log(\epsilon))$
- Accelerated gradient descent converges at a rate $O(\sqrt{\kappa} \log(\epsilon))$

Gradient-based methods

Stochastic gradient method

$$\underset{\theta}{\text{minimize}} \quad f(\theta) := \sum_{i=1}^{m} f_i(\theta), \qquad f_i \text{ is differentiable everywhere}$$

- **Stochastic gradient <u>method</u> (SGD)**

$$\theta^{(k+1)} = \theta^{(k)} - \alpha^{(k)} \nabla f_i(\theta^{(k)}), \quad i \sim \text{Unif}\{1, ..., m\}$$

- This is not a descent method
  - $f^{(k+1)}$ is not necessarily less than $f^{(k)}$, even if step size $\to 0$)
- Step size choice
  - $\alpha^{(k)}$ constant, gets to a noisy neighborhood
  - $\alpha^{(k)}$ decaying ($1/\sqrt{k}$ or $1/k$), provably convergent to local minimum

Minibatching

$$\underset{\theta}{\text{minimize}} \quad f(\theta) := \sum_{i=1}^{m} f_i(\theta), \qquad f_i \text{ is differentiable everywhere}$$

Minibatching

- Uniformly without replacement, pick a subset $\mathcal{S} \subset \{1, ..., m\}$

$$\theta^{(k+1)} = \theta^{(k)} - \alpha^{(k)} \sum_{i \in \mathcal{S}} \nabla f_i(\theta^{(k)}),$$

- Larger $\mathcal{S}$ = smaller gradient variance, greater per-iteration complexity
- When $\mathcal{S} = \{1, ..., m\}$, $\rightarrow$ gradient descent

Projections

$$\underset{\theta \in \mathcal{S}}{\text{minimize}}\ f(\theta) \qquad f \text{ convex function, } \mathcal{S} \text{ convex set}$$

- The Euclidean projection on a convex set is defined as

$$\mathbf{proj}_{\mathcal{S}}(\hat{\theta}) = \underset{\theta \in \mathcal{S}}{\operatorname{argmin}}\ \|\theta - \hat{\theta}\|_2$$

- Solution always exists and is unique when $\mathcal{S}$ is convex
- If $\hat{\theta} \in \mathcal{S}$, then $\mathbf{proj}_{\mathcal{S}}(\hat{\theta}) = \hat{\theta}$
- Some projections super easy: $\bar{\theta} = \mathbf{proj}_{\mathcal{S}}(\hat{\theta})$
    - $\mathcal{S} = \{\theta : b_k \le \theta_k \le c_k\}, \quad \bar{\theta}_k = \min\{\max\{\hat{\theta}_k, b_k\}, c_k\}$
    - $\mathcal{S} = \{\theta : \|\theta\|_2 \le 1\}, \quad \bar{\theta}_k = \frac{1}{\max\{1, \|\hat{\theta}_k\|_2\}}\hat{\theta}_k$

Projected gradient descent

$$\underset{\theta \in \mathcal{S}}{\text{minimize}}\ f(\theta) \qquad f \text{ convex function, } \mathcal{S} \text{ convex set}$$

- Projected gradient descent:

$$\theta^{(k+1)} = \mathbf{proj}_{\mathcal{S}}(\theta^{(k)} - \alpha \nabla f(\theta^{(k)}))$$

- Is a descent method
- Constant step size α is fine (if small enough)
- Basically same analysis as unprojected gradient descent

Summary

- Definition of convex set, function
  - 1st order, 2nd order conditions
- Gradient descent
  - L-smoothness, μ-strong convexity
  - Descent lemma
- Gradient-based methods
  - Stochastic GD
  - Projected GD