Day 17 - Convex Optimization and Convergence of Gradient Descent

Outline:

Convex Optimization

Convergence of GD

Optimization and machine learning

Data $\{(x_i, y_i)\}_{i=1 \cdots n}$

Consider a model $\hat{y}_\theta(x_i)$

$$\min_\theta \sum_{i=1}^{n} \ell(\hat{y}_\theta(x_i), y_i)$$

Optimization in general

$$\min_x f(x)$$

Gradient descent: Take successive steps "downhill"

$$x^{(i+1)} = x^{(i)} - \alpha \nabla f(x^{(i)})$$

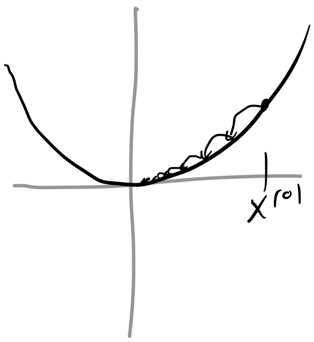
iteration index

step size, learning rate

$-\nabla f$ points in direction of steepest descent

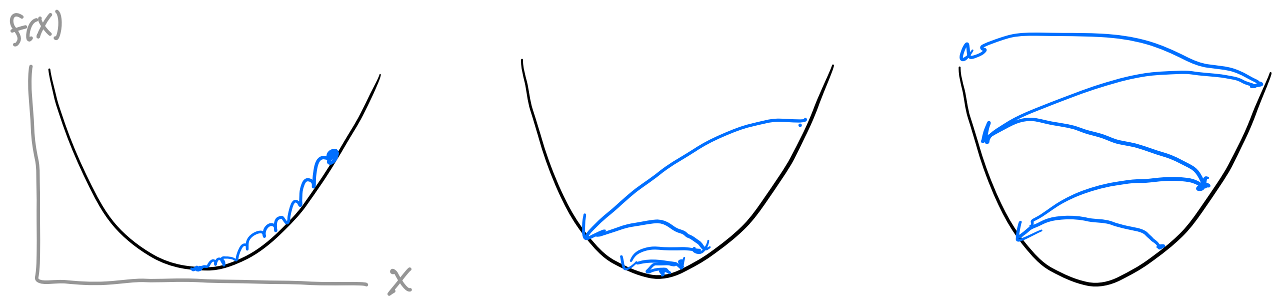
$f: \mathbb{R} \to \mathbb{R}$

Example: $f(x) = \frac{1}{2} L x^2$.



If $\alpha < \frac{2}{L}$, GD converges.

Picture:



Small learning rate	medium learning rate	high learning rate
$\alpha < \frac{1}{L}$	$\frac{1}{L} < \alpha < \frac{2}{L}$	$\alpha > 2/L$

Challenges of gradient descent in machine learning & minibatches

$$\min_{\theta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}_\theta(x_i), y_i)}_{f(\theta)}$$

$$\theta^{k+1} = \theta^k - \alpha \nabla f(\theta) = \theta^k - \alpha \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(\hat{y}_\theta(x_i), y_i)$$

To evaluate $\nabla f(\theta)$, one needs to loop through all data (batch gradient descent)

- expensive
- not possible in some contexts

Idea: use minibatches

Select a minibatch $B \subset \{1, 2, \dots, n\}$

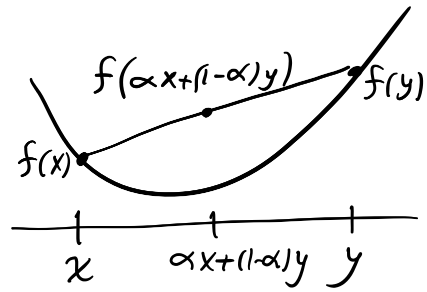
$$\theta^{k+1} = \theta^k - \alpha \underbrace{\frac{1}{|B|}\sum_{i \in B} \nabla_\theta \ell(\hat{y}_\theta(x_i), y_i)}_{\text{use as approximation of } \nabla_\theta f(\theta)}$$

Convex Optimization

We say $f: \mathbb{R}^d \to \mathbb{R}$ is convex if

$$f(\underbrace{\alpha x + (1-\alpha) y}_{\text{Convex combo}}) \leq \alpha f(x) + (1-\alpha) f(y)$$

for all $0 \leq \alpha \leq 1$, x, y.

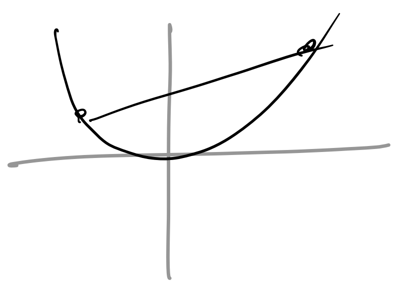


"always curves up"

Examples:

$f: \mathbb{R} \to \mathbb{R}$

$f(x) = x^2$ is or is not convex



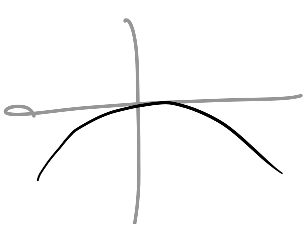
Fix a $c \in \mathbb{R}$.

$f: \mathbb{R} \to \mathbb{R}$

$f(x) = c x^2$ is or is not convex

If $c \geq 0$, yes

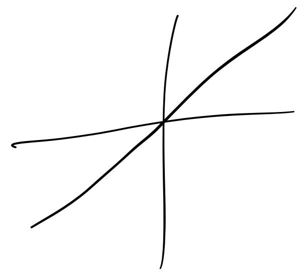
$c < 0$, no



$f: \mathbb{R} \to \mathbb{R}$

$f(x) = x$

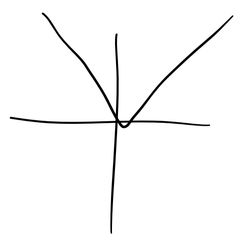
is or is not convex



$f: \mathbb{R} \to \mathbb{R}$

$f(x) = |x|$

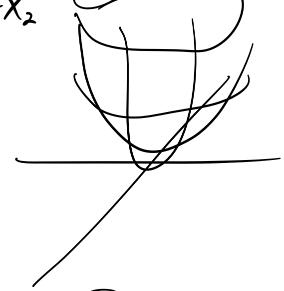
is or is not convex



$f: \mathbb{R}^2 \to \mathbb{R}$

$f(x) = \|x\|^2 = x_1^2 + x_2^2$

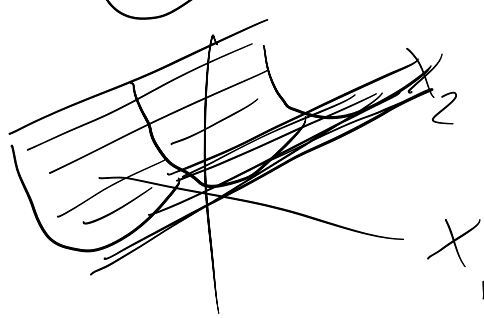
is or is not convex



$f: \mathbb{R}^2 \to \mathbb{R}$

$f(x) = x_1^2$

is or is not convex



We will study the minimization of convex functions.

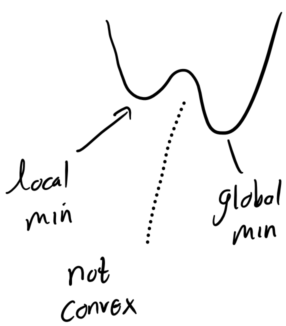
Does every convex function f have a minimal value?

$$\min_x f(x)$$

No: $f(x) = x$

or $= e^x$

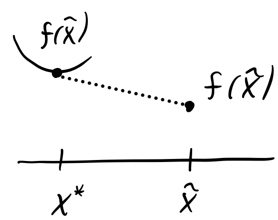
All local minima of convex functions are global minima.



Suppose x_* is a local min of f.

If $x \approx x_*$ the $f(x) \geq f(x_*)$.

Suppose x_* is not a global min.



Suppose $f(\hat{x}) < f(x^*)$.

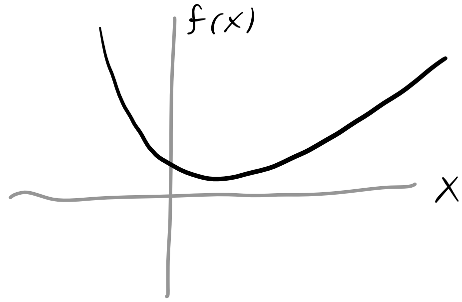
By convexity, $f(x)$ lies below dotted line between x^* and $\hat{x}$. So x^* not a local min.

Contradiction. $\times$

Convexity and Second derivatives

Functions of one variable

If $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable everywhere, f is convex if and only if $f''(x) \geq 0$ for all x.



Functions of multiple variables

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable

f is convex if $D^2 f = Hf$ is positive semidefinite everywhere

Hessian matrix

$$D^2 f = Hf(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \vdots & & \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

H is positive definite if all eigenvalues are positive

H is positive semidefinite if all eigenvalues are nonnegative

Eigenvalue Decomposition:

If $H \in \mathbb{R}^{n \times n}$ is symmetric ($H^t = H$), then H has an orthonormal basis of eigenvectors with real eigenvalues. So

$$H = \underbrace{U}_{n \times n} \underbrace{\Lambda}_{\substack{\text{diagonal} \\ n \times n}} U^t$$

where U has orthonormal columns

could have negative entries

We say v_i is an eigenvector of H with eigenvalue λ_i if $Hv_i = \lambda_i v_i$

$$H = \underbrace{\begin{pmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{pmatrix}}_{U} \cdot \underbrace{\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}}_{\Lambda} \cdot \underbrace{\begin{pmatrix} - & v_1^t & - \\ - & v_2^t & - \\ & \vdots & \\ - & v_n^t & - \end{pmatrix}}_{U^t}$$

Columns are unit length eigenvectors that are orthogonal to each other

$$v_i \cdot v_j = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$$

We also have

$$H = \sum_{i=1}^{n} \lambda_i v_i v_i^t .$$

Why?

$$H = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \begin{pmatrix} - & v_1^t & - \\ & \vdots & \\ - & v_n^t & - \end{pmatrix}$$

$$= \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix} \begin{pmatrix} - & \lambda_1 v_1^t & - \\ & \vdots & \\ - & \lambda_n v_n^t & - \end{pmatrix} = \sum_{i=1}^{n} v_i (\lambda_i v_i^t)$$

Theorem: H is positive semidefinite if and only if

$$z^t H z \geq 0 \quad \text{for all } z \in \mathbb{R}^n$$

Recall, because H is symmetric ($H = H^t$), H has an orthonormal basis of eigenvectors with real eigenvalues. So

$$H = U \Lambda U^t$$

where U has orthonormal columns

and Λ is diagonal

Proof of Theorem: • PSD $\Rightarrow$ $z^t H z \geq 0$ for all z

As H is PSD, Λ has nonneg. diagonal entries. So $z^t H z = z^t U \Lambda U^t z$

$$= \sum_{i=1}^{n} \Lambda_{ii} (U^t z)_i^2$$

$$\geq 0.$$

• $z^t H z \geq 0$ for all z $\Rightarrow$ H is PSD

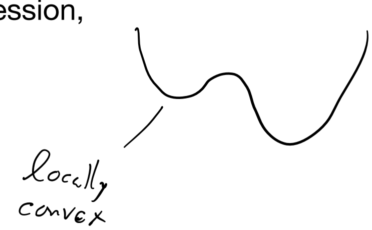
Suppose H is not PSD. At least one eigenvalue is negative. Suppose U_i is eigenvector w/ e-val $\lambda_i < 0$. Then let $z = U_i$.

$$z^t H z = U_i^t H U_i = \lambda_i U_i^t U_i < 0.$$

Many but not all ML optimization problems are convex.

Convex Problems: least squares regression, logistic regression,

Not convex: neural networks



How fast does gradient descent converge?

$$\min_{x} f(x), \qquad x^{(i+1)} = x^{(i)} - \alpha \nabla f(x^{(i)})$$

Suppose $x^{(i)} \to x^*$ as $i \to \infty$.

How long do you need to wait to get a certain accuracy ε?

Can gain understanding in some convex cases.

Convergence of GD for quadratic functions

Let $f(x) = \frac{1}{2} x^t Q x - b^t x$

where $x \in \mathbb{R}^d$, $b \in \mathbb{R}^d$, $Q \in \mathbb{R}^{d \times d}$ is positive definite

Let $m = \lambda_{min}(Q)$, $M = \lambda_{max}(Q)$, $K = \frac{M}{m}$

condition number of Q

Consider GD w/ fixed step size α

$$x^{k+1} = x^k - \alpha \nabla f(x^k)$$

Note: $x^* = Q^{-1} b$ is the unique global min of f

Analytically show that this is the solution to the problem

$$\nabla f(x) = Qx - b = 0$$

$$Qx = b \Rightarrow x = Q^{-1} b$$

Theorem: If $\alpha = \frac{2}{M+m}$, then GD

for $f(x) = \frac{1}{2} x^t Q x - b^t x$ satisfies

$$\|x^k - x^*\| \le \left(\frac{1-1/K}{1+1/K}\right)^k \|x^0 - x^*\|$$

"first-order convergence"

Error decays exponentially

To get error ε, need $O(\log(\varepsilon^{-1}))$ iterations

Proof: Note $\nabla f(x) = Qx - b$.

The global minimizer solves $Qx^* = b \Rightarrow x^* = Q^{-1}b$

$$\begin{aligned} x^{k+1} - x^* &= x^k - \alpha \nabla f(x^k) - x^* \\ &= x^k - \alpha(Qx^k - b) - x^* \\ &= x^k - \alpha(Qx^k - Qx^*) - x^* \\ &= (I - \alpha Q)(x^k - x^*) \end{aligned}$$

So,

$$\|x^{k+1} - x^*\| \le \underbrace{\|I - \alpha Q\|}_{= \max(\alpha M - 1,\, 1 - \alpha m)} \|x^k - x^*\|$$

largest eigenvalue of $I - \alpha Q$ (in terms of abs value)

We choose $\alpha = \dfrac{2}{M+m}$.

So $\|I - \alpha Q\| = \dfrac{M-m}{M+m} = \dfrac{1-1/k}{1+1/k} < 1$

$$\Rightarrow \|x^{k+1} - x^*\| \leq \left(\frac{1-1/k}{1+1/k}\right) \|x^k - x^*\|$$

$$\Rightarrow \|x^k - x^*\| \leq \left(\frac{1-1/k}{1+1/k}\right)^k \|x^0 - x^*\| \quad \blacksquare$$

Interpretation:

If f doesnt curve up too much and doesnt curve up too little, then GD with fixed step size can exhibit first order convergence to the global minimizer

Should we think of GD as converging "quickly"?

If the function is quadratic, then GD (with the right step size) can converge very quickly.

If the function is not quadratic, then GD may converge slowly

Theorem: Let f be convex and $\lambda_{max}(Hf(x)) \leq M$ for all x. If $\alpha \leq \frac{1}{M}$, then GD satisfies

$$f(x^{(i)}) - f(x^*) \leq \frac{1}{2i\alpha} \|x^{(0)} - x^*\|^2$$

where x^* is a minimizer of f.

- Error decays slowly
- To get error ε from optimal value, need $O(\varepsilon^{-1})$ iterations

Summary:

- Too large learning rate can lead to divergence
- In convex case, to get convergence α should be small relative to curvature of f
- Too small learning rate can lead to slow convergence
- For convex quadratic functions, convergence of GD can be first order (fast)
- For more general convex functions, convergence can be slow
- SGD w/ fixed step size is not expected to converge
- SGD with decaying step sizes may converge