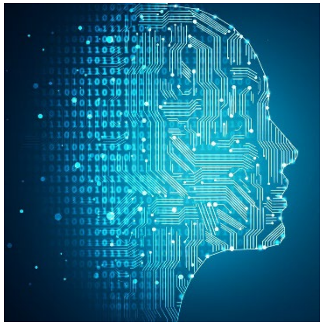
Machine Learning

Unconstrained Optimization

Kevin Moon (kevin.moon@usu.edu)

STAT 6655



Outline

1. Minimums and necessary conditions
2. Convexity
3. Methods for solving optimization problems

Optimization in machine learning

- Many (all?) machine learning problems can be posed as a minimization or maximization problem
  - Empirical risk minimization
- In some cases a closed form solution exists
  - E.g. linear regression
- Most of the time, a closed form solution doesn't exist
- We can solve these problems using optimization theory

Unconstrained Optimization

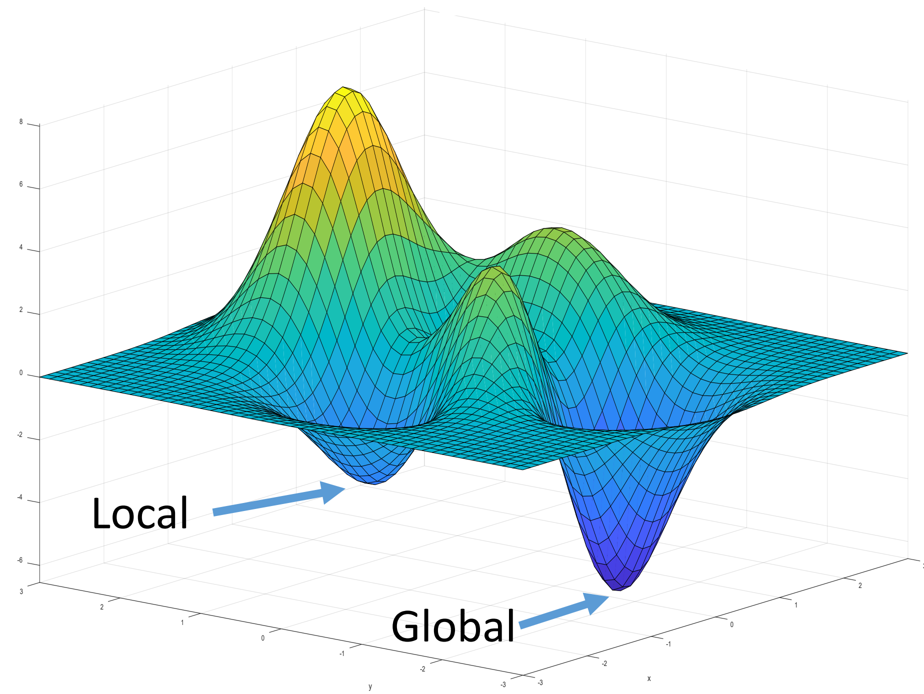
- An *unconstrained optimization problem* has the form

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

$\exists$="there exists"

$\forall$="for all"

where $f : \mathbb{R}^d \to \mathbb{R}$ is called the *objective function*.
  - What about maximization?
- A point $\boldsymbol{x}^* \in \mathbb{R}^d$ is called a *local minimizer* if $\exists r > 0$ such that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) \forall \boldsymbol{x}$ satisfying $||\boldsymbol{x} - \boldsymbol{x}^*|| < r$
- x^* is called a *global minimizer* if $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) \ \forall \boldsymbol{x} \in \mathbb{R}^d$
- Is a global minimizer also a local minimizer?

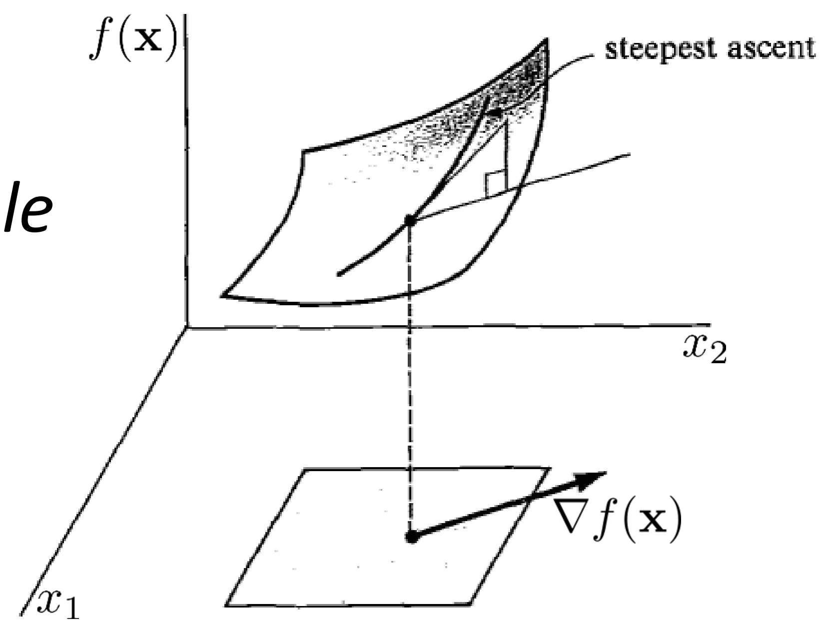


Gradient

- Given a function $f: \mathbb{R}^d \to \mathbb{R}$, the *gradient* of f at $\boldsymbol{x} = [x_1 \ldots x_d]^T \in \mathbb{R}^d$ is defined by

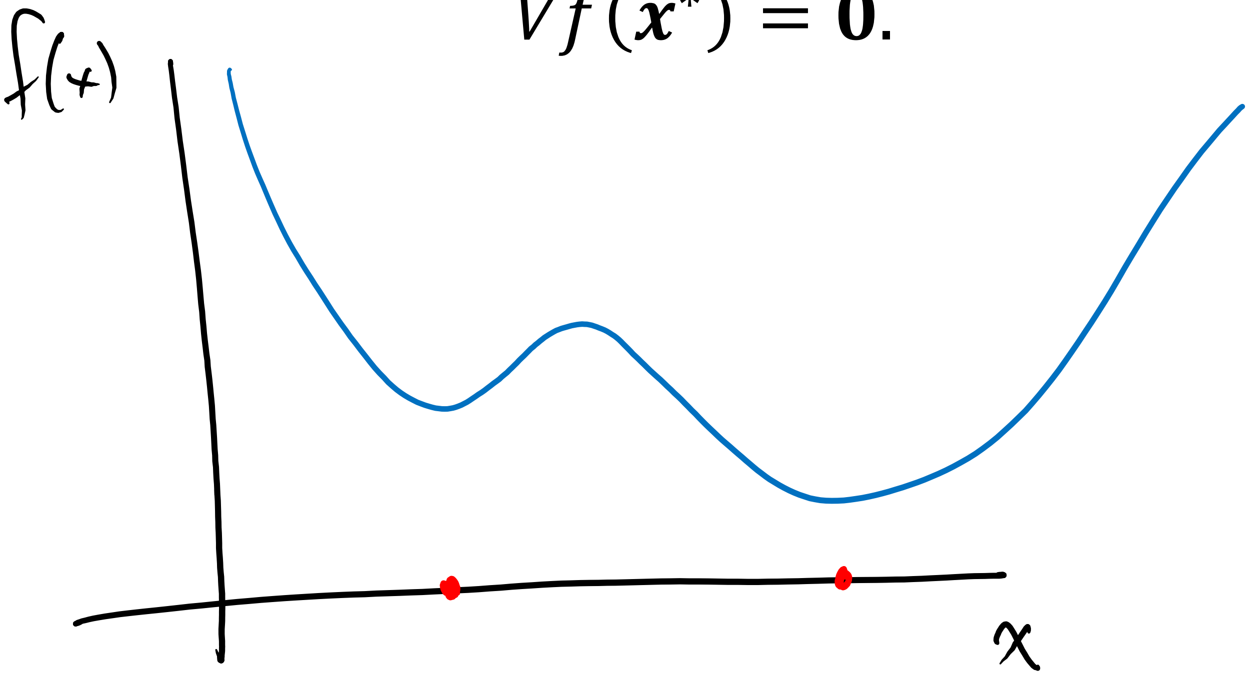
$$\nabla f(\boldsymbol{x}) := \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_d} \end{bmatrix}$$

- If the gradient exists for all $\boldsymbol{x} \in \mathbb{R}^d$, we say f is *differentiable*
- The gradient gives the direction of *steepest ascent*



First Order Necessary Condition

- If f is differentiable and $\boldsymbol{x}^*$ is a local minimizer of f, then
$$\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}.$$



- Note that this condition is necessary but <u>not</u> sufficient for $\boldsymbol{x}^*$ to be a local minimizer
  - Why?
- If $\nabla f(\boldsymbol{x}) = 0$ for some $\boldsymbol{x}$, then $\boldsymbol{x}$ is said to be a critical point or a stationary point of f

First Order Necessary Condition

- *Proof:* Define the scalar valued function $\phi(t) = f(\boldsymbol{x}^* + \boldsymbol{y}t)$, where $\boldsymbol{y} \in \mathbb{R}^d$ is arbitrary. Then

$$\phi'(0) = \lim_{t \searrow 0} \frac{f(\boldsymbol{x}^* + \boldsymbol{y}t) - f(\boldsymbol{x}^*)}{t}$$
$$= \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} \rangle$$

by the chain rule. Since $\boldsymbol{x}^*$ is a local min, we know

$$f(\boldsymbol{x}^* + \boldsymbol{y}t) \geq f(\boldsymbol{x}^*)$$

for t sufficiently small. Therefore, $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} \rangle \geq 0$. Now choose $\boldsymbol{y} = -\nabla f(\boldsymbol{x}^*)$. Then

$$0 \leq \langle \nabla f(\boldsymbol{x}^*), -\nabla f(\boldsymbol{x}^*) \rangle = -\|\nabla f(\boldsymbol{x})\|^2 \leq 0,$$

so we must have $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

Hessian

- The *Hessian* of f at $\boldsymbol{x}$ is the $d \times d$ matrix

$$\nabla^2 f(\boldsymbol{x}) := \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_d^2} \end{bmatrix}$$

- We say that f is *twice differentiable* if $\nabla^2 f(\boldsymbol{x})$ exists $\forall \boldsymbol{x} \in \mathbb{R}^d$.
- We say f is *twice continuously differentiable* if it is twice differentiable and all of the second derivatives are continuous functions of $\boldsymbol{x}$.
- If f is twice continuously differentiable, then $\nabla^2 f(\boldsymbol{x})$ is a symmetric matrix $\forall \boldsymbol{x}$, i.e.,

$$\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_j \partial x_i}$$
$$\forall \boldsymbol{x} \in \mathbb{R}^d$$
$$\forall i, j = 1, \ldots, d$$

Positive (Semi-)Definite Matrices

- Let A be a $d \times d$ matrix. We say that A is *positive definite* (PD) if $\boldsymbol{x}^T A \boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$. We say that A is *positive semi-definite* (PSD) if $\boldsymbol{x}^T A \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$.
- PD and PSD arise frequently in ML, for example
  - Gram matrices
  - Kernel matrices
  - Covariance matrices
  - Hessian matrices (sometimes)
- PD/PSD matrices are not necessarily symmetric, e.g.

$$\begin{bmatrix} x \\ y \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x & y \end{bmatrix} = x^2 + y^2$$

  - However, we will only consider PD/PSD matrices that are also symmetric

Second Order Necessary Condition

- If f is twice continuously differentiable and $\boldsymbol{x}^*$ is a local min, then $\nabla^2 f(\boldsymbol{x}^*)$ is positive semi-definite, i.e.,

$$\boldsymbol{z}^T \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{z} \geq 0, \qquad \forall \boldsymbol{z} \in \mathbb{R}^d$$

  - This generalizes the result from single-variable calculus that the second derivative is nonnegative at a local min

- *Proof:*
  From the definition of local optimality, there exists a neighborhood A of $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*)$ is the minimum inside A.
  - I.e., $\exists r > 0$ such that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) \forall \boldsymbol{x}$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}^*\| < r$ and $A = \{\boldsymbol{x} \in \mathbb{R}^d \big| \|\boldsymbol{x} - \boldsymbol{x}^*\| < r\}$

Proof continued

- By multidimensional Taylor series expansion, we can write for any $\boldsymbol{y}$ and t such that $\boldsymbol{x}^* + t\boldsymbol{y} \in A$:

$$f(\boldsymbol{x}^* + t\boldsymbol{y})$$
$$= f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*), t\boldsymbol{y} \rangle + \frac{t^2}{2}\left\langle \boldsymbol{y}, \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{y} \right\rangle + o\left(t^2 \|\boldsymbol{y}\|^2\right)$$

  - $o(t)$ denotes a function satisfying $\lim_{t \to 0} \frac{o(t)}{t} = 0$

- Noting that $\nabla f(\boldsymbol{x}^*) = 0$ and rearranging gives for $y \neq 0$:

$$0 \leq \frac{f(\boldsymbol{x}^* + t\boldsymbol{y}) - f(\boldsymbol{x}^*)}{t^2 \|\boldsymbol{y}\|^2} = \frac{o\left(t^2 \|\boldsymbol{y}\|^2\right)}{t^2 \|\boldsymbol{y}\|^2} + \frac{1}{2\|\boldsymbol{y}\|^2}\left\langle \boldsymbol{y}, \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{y} \right\rangle$$

  - The inequality follows from the local optimality of $\boldsymbol{x}^*$

- Taking the limit of $t \to 0$ of both sides gives

$$\frac{1}{2\|\boldsymbol{y}\|^2}\left\langle \boldsymbol{y}, \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{y} \right\rangle \geq 0$$

  - This statement proves that the Hessian is PSD

Group Exercise

Notation:

$$\boldsymbol{x} = \begin{bmatrix} x \\ y \end{bmatrix}$$

Consider the function

$$f(x, y) = x^2 + 4xy - y^2 - 8x - 6y + 10$$

1. Determine $\nabla f(x, y)$
2. Determine $\nabla^2 f(x, y)$
3. Determine a critical point $\boldsymbol{x}^*$
4. Is $\boldsymbol{x}^*$ a local min, a local max, or neither?

Convexity

- We say that f is *convex* if

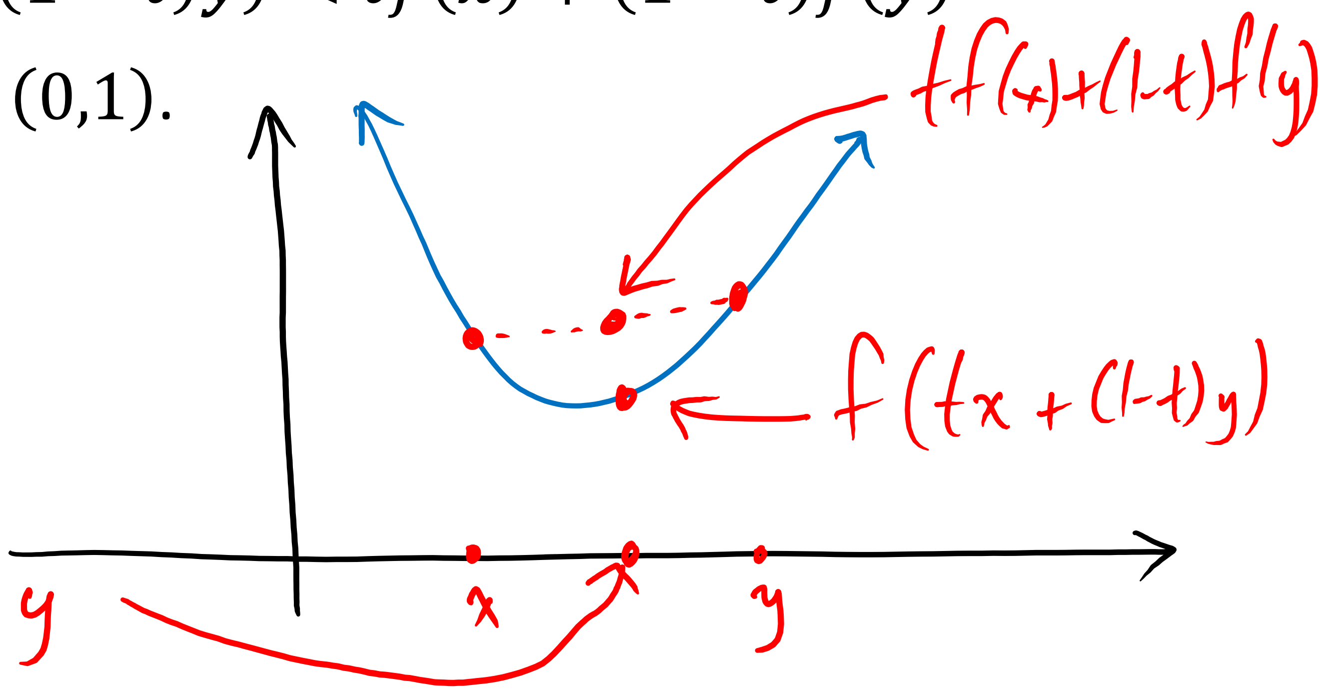
$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y})$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $t \in [0,1]$.

- We say f is *strictly convex* if

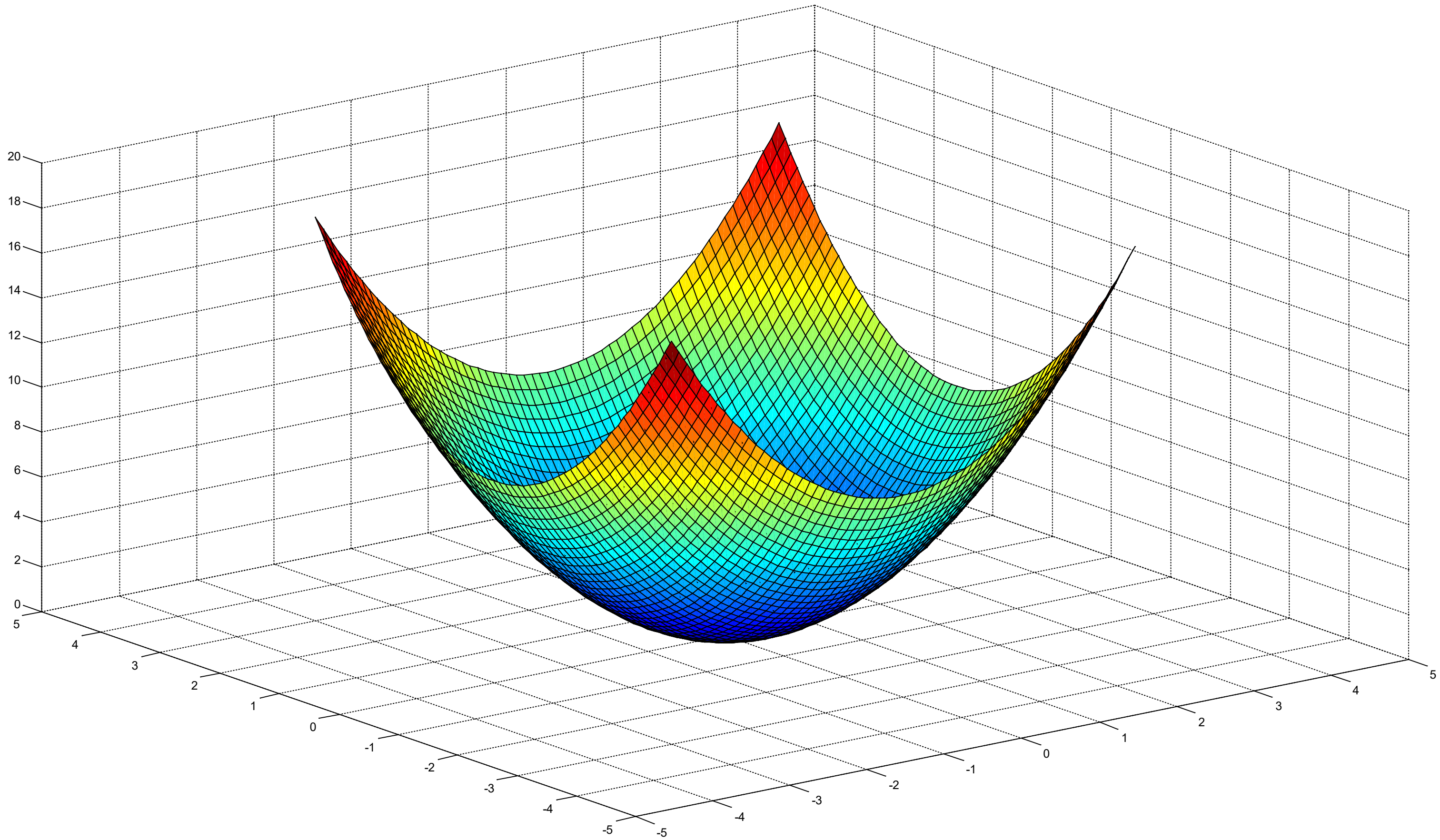
$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) < tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y})$$

$\forall \boldsymbol{x} \neq \boldsymbol{y}$ and $t \in (0,1)$.



Convexity

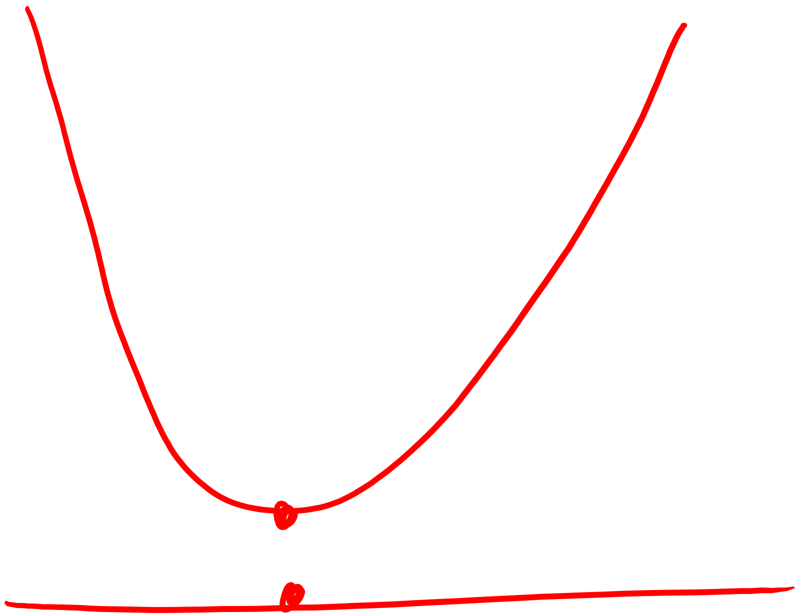
$$f(x, y) = x^2 + y^2$$



Convex functions are nice

Properties of convex functions

1. If f is convex, then every local min is a global min
2. If f is strictly convex, then f has at most one global min



Convex functions are nice

Proof of 1: Suppose $\boldsymbol{x}^*$ is a local min but not a global min. Then $\exists \boldsymbol{y}^* \in \mathbb{R}$ such that $f(\boldsymbol{y}^*) < f(\boldsymbol{x}^*)$. By convexity, $\forall t \in [0,1)$ we have

$$\begin{aligned} f(t\boldsymbol{x}^* + (1-t)\boldsymbol{y}^*) &\leq tf(\boldsymbol{x}^*) + (1-t)f(\boldsymbol{y}^*) \\ &< tf(\boldsymbol{x}^*) + (1-t)f(\boldsymbol{x}^*) \\ &= f(\boldsymbol{x}^*) \end{aligned}$$

Taking $t \nearrow 1$, the above strict inequality contradicts local minimality of $\boldsymbol{x}^*$. Thus $\boldsymbol{x}^*$ is a global min.

Group Exercise

1. Give an example of a function f that is
   - convex but not strictly convex
   - convex and has more than one global minimum
   - strictly convex, but has no global minimum
2. Is the sum of convex functions necessarily convex? Prove or provide a counter example
3. Is the product of convex functions necessarily convex? Prove or provide a counter example

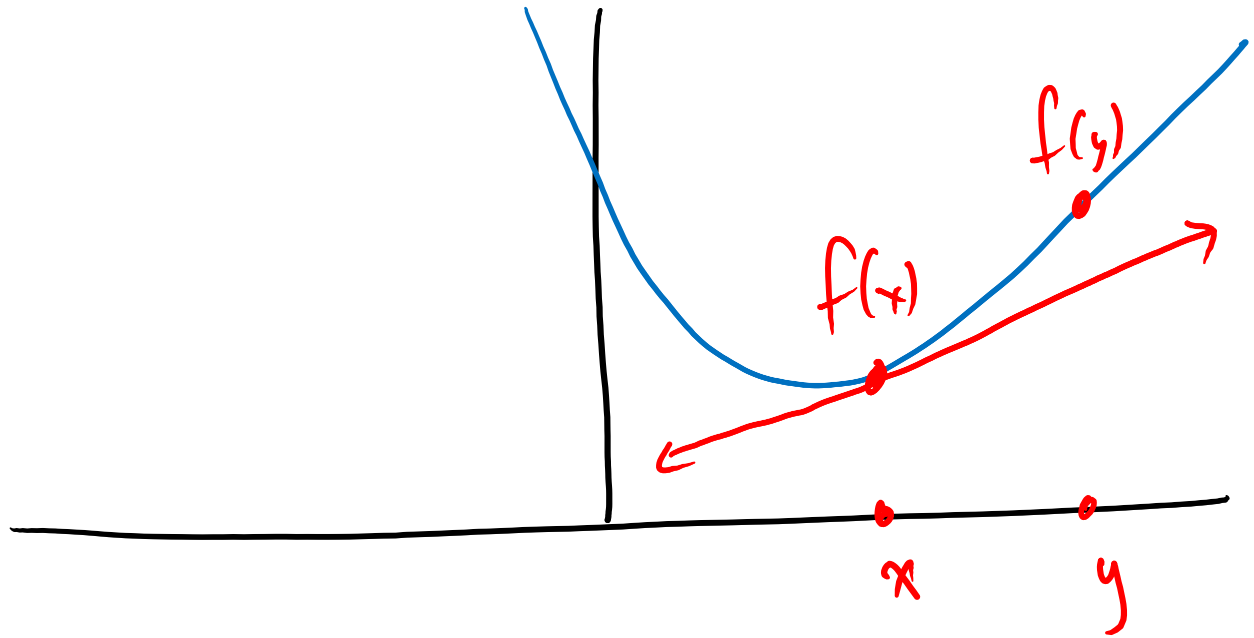
1^{st} order characterization of convexity

1. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Then f is convex iff $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle.$$

2. Similarly, f is strictly convex iff $\forall \boldsymbol{x} \neq \boldsymbol{y}$,

$$f(\boldsymbol{y}) > f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle.$$



1st order characterization of convexity

Proof of 1 ($\Longrightarrow$): First, assume $\boldsymbol{x}$ is convex. For any $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d, t \in [0,1]$,

$$\begin{aligned} f(t\boldsymbol{y} + (1-t)\boldsymbol{x}) &\leq tf(\boldsymbol{y}) + (1-t)f(\boldsymbol{x}) \\ &= f(\boldsymbol{x}) + t(f(\boldsymbol{y}) - f(\boldsymbol{x})). \end{aligned} \tag{1}$$

Rearranging,

$$\frac{f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x})}{t} \leq f(\boldsymbol{y}) - f(\boldsymbol{x})$$

The limit of the LHS is a directional derivative and equal to $\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$ by the chain rule. Therefore $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$.

1^{st} order characterization of convexity

Proof of 1 ($\Longleftarrow$): Now suppose conversely that $\forall \boldsymbol{x}, \boldsymbol{y}$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \qquad *$$

Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $t \in [0, 1]$. Denote $\boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}$. Applying * twice we have

$$f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{x} - \boldsymbol{z} \rangle \qquad * \, a$$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{y} - \boldsymbol{z} \rangle \qquad * \, b$$

Now multiply $*a$ by t, $*b$ by $(1-t)$ and add:

$$\begin{aligned} tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) &\geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), t\boldsymbol{x} + (1-t)\boldsymbol{y} - \boldsymbol{z} \rangle \\ &= f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \end{aligned}$$

as $t\boldsymbol{x} + (1-t)\boldsymbol{y} - \boldsymbol{z} = 0$. This establishes convexity.

1st order condition for local min, revisited

- Let f be convex and continuously differentiable. Then $\boldsymbol{x}^*$ is a global min iff $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.
- Proof:
  (⇒) Already discussed

$$(\Leftarrow) \quad \begin{aligned} \forall \boldsymbol{y}, f(\boldsymbol{y}) &\geq f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x} \rangle \\ &= f(\boldsymbol{x}^*) \end{aligned}$$

- Thus for convex functions, $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ is both necessary and sufficient for $\boldsymbol{x}^*$ to be a global min.

Second order conditions of convexity

More important properties:

- f is convex $\Leftrightarrow \nabla^2 f(\boldsymbol{x})$ is positive semidefinite $\forall \boldsymbol{x} \in \mathbb{R}^d$
- f is strictly convex $\Leftarrow \nabla^2 f(\boldsymbol{x})$ is positive definite $\forall \boldsymbol{x} \in \mathbb{R}^d$

- Proofs follow from multidimensional Taylor series expansions and the property that if f is convex, then $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$

Group Exercises

1. Give an example of a function f and a point $\boldsymbol{x}$ such that $\nabla^2 f(\boldsymbol{x})$ is PSD but $\boldsymbol{x}$ is not a local minimizer

2. Give an example of a function f that is strictly convex, but such that there exists $\boldsymbol{x}$ for which $\nabla^2 f(\boldsymbol{x})$ is not PD.

3. Numerically determine a critical point of

$$f(x, y) = x^2 + 2xy + 3y^2 + 4x + 5y + 6$$

   and also determine if it is a local/global min or max. *Note:* If you don't have access to Matlab/Python/etc. in class, you can also use Wolfram Alpha for many calculations like eigenvalue decompositions

Regularized Logistic Regression

- Unless $n \gg d$, it is preferable to minimize the modified objective function in logistic regression:

$$J(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||^2$$

  - $\lambda > 0$ is a fixed, user-specified constant called a *regularization parameter*
- Why introduce the regularization term?
  - So the Hessian is PD (see a future HW)

Methods for solving optimization problems

1. Gradient Descent
2. Stochastic Gradient Descent
3. Newton’s method
4. Subgradient methods

Iterative methods

- In many problems, an analytical solution to the minimization problem does not exist
  - Including (regularized) logistic regression
- However, we can use an iterative method to solve it
  - Convexity guarantees that the iterative solution is the global minimum

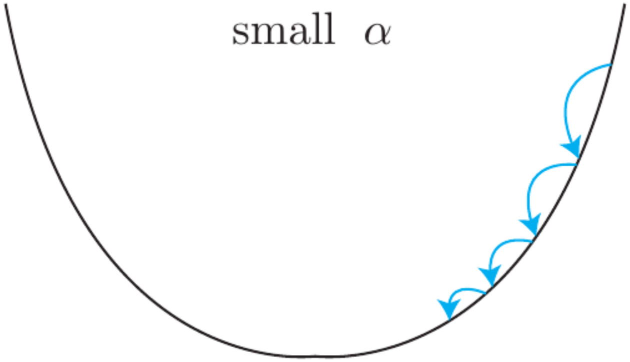
Gradient Descent (GD)

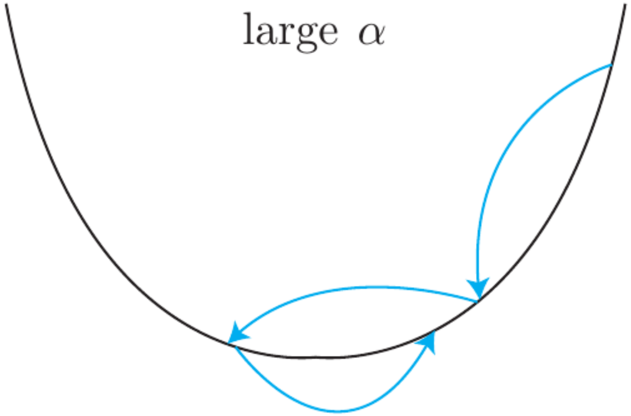
- Consider minimizing the generic objective function $J(\boldsymbol{\theta})$
- What is a geometric interpretation of $\nabla J(\boldsymbol{\theta})$?
  - Direction of steepest ascent
  - Thus iterative approaches take steps in opposite direction, i.e., the direction of steepest descent
- Initial guess $\boldsymbol{\theta}_0$
- For $t = 1, \ldots, \mathrm{max\_iter}$

  $$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha \nabla J(\boldsymbol{\theta}_{t-1})$$

  If convergence condition satisfied, exit

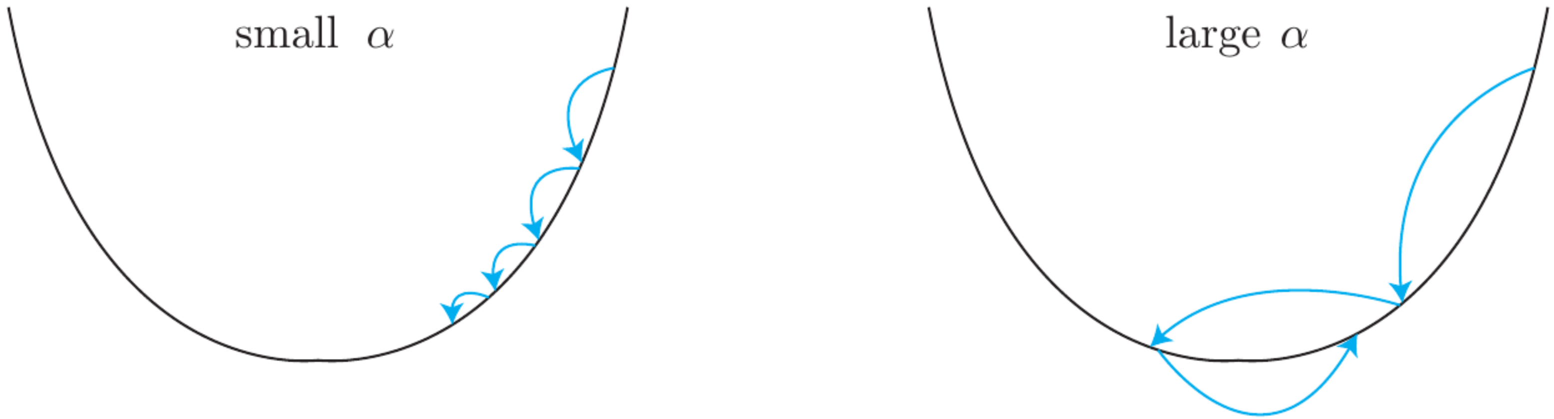
  End





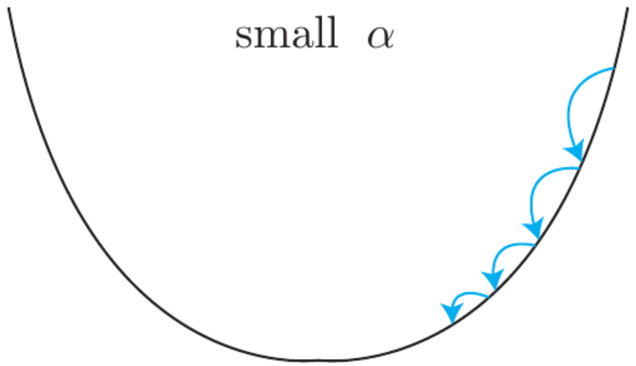
Step Size

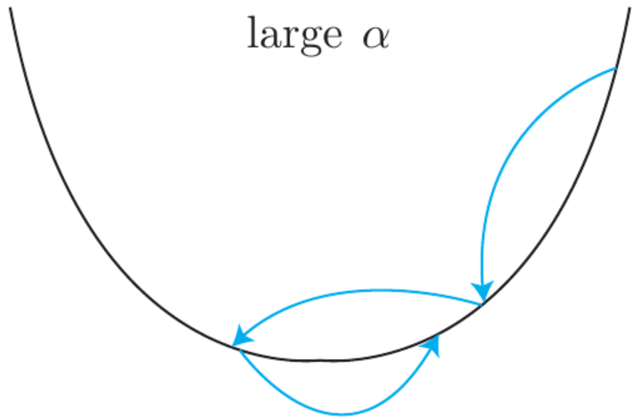
- In practice, typically need α to decrease as a function of t
  - Example: $\alpha_t = \frac{1}{t}$



Group Exercise

1. What would be a good convergence condition (also known as a stopping criterion) for terminating GD?
2. What are some potential problems with setting $\alpha_t = \frac{1}{t}$?
3. How could you choose α_t to counter these problems?
4. Does the initialization $\boldsymbol{\theta}_0$ affect the result?





Stochastic Gradient Descent (SGD)

- In many ML problems, we can write

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} J_i(\boldsymbol{\theta})$$

  - For example, in regularized ERM, we can write

$$J_i(\boldsymbol{\theta}) = \frac{1}{n}\Big(\mathrm{L}\big(\mathrm{y_i}, \mathrm{f}_{\boldsymbol{\theta}}(\mathbf{x}_\mathrm{i})\big) + \lambda\Omega(f_{\boldsymbol{\theta}})\Big)$$

- Thus

$$\nabla J(\boldsymbol{\theta}) = \sum_{i=1}^{n} \nabla J_i(\boldsymbol{\theta})$$

Stochastic Gradient Descent (SGD)

- In GD, we calculate the gradient using all of the data
  - For large n, each step in the parameter space $\boldsymbol{\theta}$ takes a lot of computations
  - Thus learning can occur slowly
- For SGD, we estimate the gradient using a small random sample of training inputs
  - Speeds up gradient descent and learning

Stochastic Gradient Descent

- Pick a random set of $m < n$ training inputs
  - Referred to as a *mini-batch*
- Estimate the gradient using the samples from the mini-batch and take a step in the direction of the negative gradient
- If m is large enough, then the estimated gradient will be roughly equal to the gradient using the full data

Stochastic Gradient Descent Summary

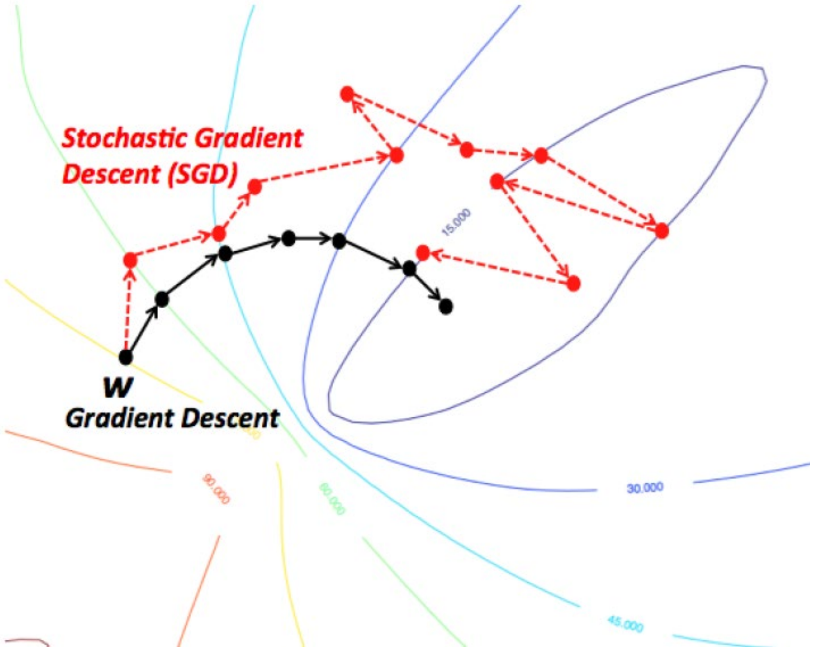
A training ***epoch***:

1. Pick a random subset of the training data
   - Referred to as a mini-batch
2. Update the parameters using the gradient estimates from the mini-batch
3. Pick another random mini-batch from the remaining training points and repeat step 2
   - Repeat until all training inputs have been used

Repeat multiple epochs until stopping conditions are met

Analogy to political polling

- Much easier to carry out a poll than run a full election
- Similarly, it's much easier to estimate gradients from mini-batches than the entire training set
- Downside: gradient estimates will be noisier in SGD
- Generally ok as we only need to move in a general direction that decreases *J*
  - This can be an advantage when the problem is nonconvex
- In practice, SGD (or its variants) are used extensively in learning neural networks



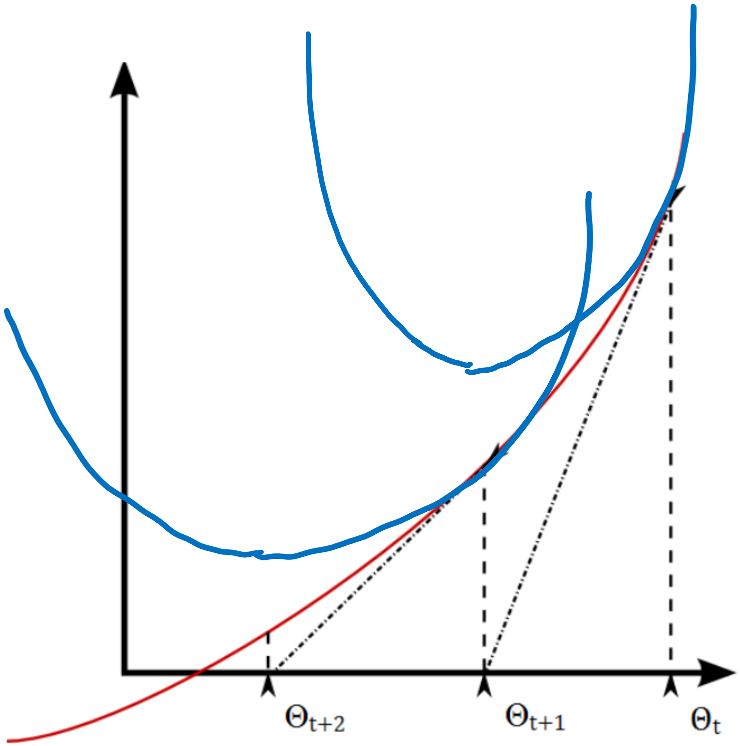
Newton's method

- Goal is to solve $\nabla J(\boldsymbol{\theta}) = \mathbf{0}$ for an objective function J
- Newton's method AKA Newton-Raphson:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \left(\nabla^2 J(\boldsymbol{\theta}_t)\right)^{-1} \nabla J(\boldsymbol{\theta}_t)$$

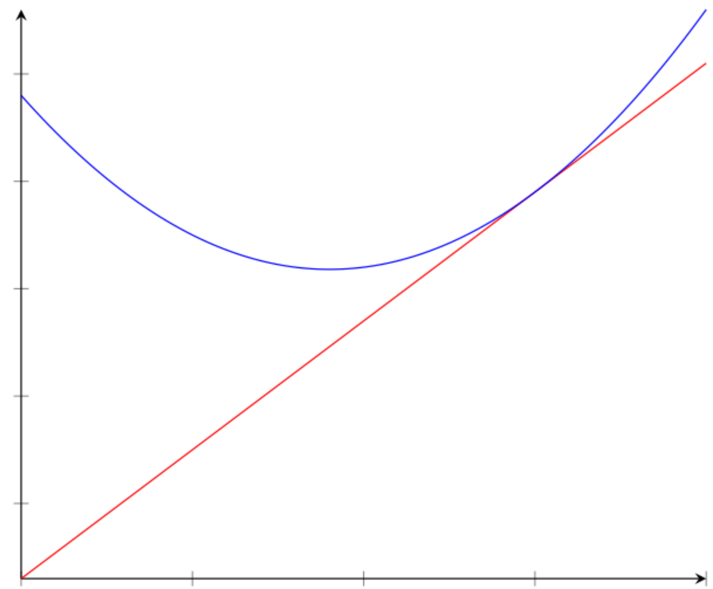
- Newton's method can be viewed as minimizing the second order approximation:

$$J(\boldsymbol{\theta}) \approx J(\boldsymbol{\theta}_t) + \nabla J(\boldsymbol{\theta}_t)^T (\boldsymbol{\theta} - \boldsymbol{\theta}_t) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_t)^T \nabla^2 J(\boldsymbol{\theta}_t) (\boldsymbol{\theta} - \boldsymbol{\theta}_t)$$



Subgradient methods

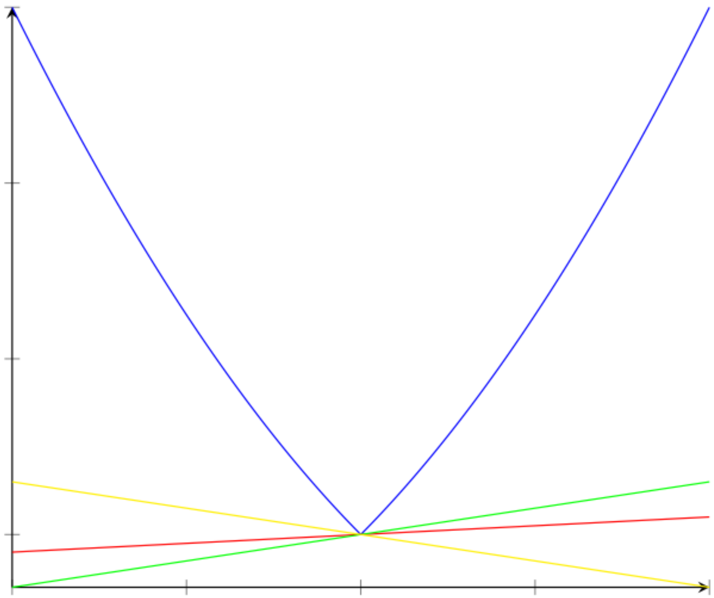
- Subgradient methods are generalizations of gradient descent that can be applied to *nondifferentiable*, *convex* objective functions, like ERM with hinge loss
  - Hinge loss: $L(y,t) = \max(0,1-yt)$
- Let $g: \mathbb{R}^d \to \mathbb{R}$ be convex and let $\boldsymbol{\theta} \in \mathbb{R}^d$. If g is differentiable, then $\boldsymbol{u} = \nabla g(\boldsymbol{\theta})$ is the only vector such that



$$g(\boldsymbol{\theta}') \le g(\boldsymbol{\theta}) + \boldsymbol{u}^T(\boldsymbol{\theta}' - \boldsymbol{\theta}) \;\; \forall \boldsymbol{\theta}'$$

Subgradients

- If g is convex but *not* differentiable, then for some $\boldsymbol{\theta}$, there may be many $\boldsymbol{u}$ satisfying the previous inequality.
- We define the *subdifferential* of g at $\boldsymbol{\theta}$, denoted $\partial g(\boldsymbol{\theta})$, to be the set of all $\boldsymbol{u}$ satisfying the inequality.
- A *subgradient* is any element of the subdifferential.
- In the figure, the subdifferential is the interval $\left[g'_{-}(\theta), g'_{+}(\theta)\right]$ where $g'_{-}(\theta), g'_{+}(\theta)$ denote the left and right derivatives.



Subgradient Methods

- In a *subgradient method*, we update the parameter just as in gradient descent, but where the gradient is replaced by *any* subgradient
- Pseudo-code for minimizing $g(\boldsymbol{\theta})$
  - Initialize $\boldsymbol{\theta}_0$
  - $t \leftarrow 0$
  - Repeat
    - Select $\boldsymbol{u}_t \in \partial g(\boldsymbol{\theta}_t)$
    - $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \alpha_t \boldsymbol{u}_t$
    - $t \leftarrow t + 1$
  - Until stopping criterion satisfied
- If we can write $g(\boldsymbol{\theta}) = \sum_{i=1}^{n} g_i(\boldsymbol{\theta})$, then we can also have a *stochastic subgradient method*, analogous to SGD.

Big Takeaways

- The combination of convexity and differentiability enable us to determine if a global minimum exists
- Differentiability also provides us with methods for finding a minimum when a closed form solution may not exist
  - If convex but not differentiable, then subgradient methods may be useful
  - Other methods also exist such as the Majorize-Minimize Algorithm
- Thus when considering loss functions for an ML problem, convexity and differentiability are important properties to consider

Further reading

- ESL Section 10.10 and 11.4