CS-GY 6763: Lecture 7
Gradient Descent and Projected Gradient Descent

NYU, Prof. Ainesh Bakshi

optimization

Next Unit: Continuous Optimization

Have some function $f : \mathbb{R}^d \to \mathbb{R}$. Want to find $\mathbf{x}^* \in R^d$ such that:

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

Next Unit: Continuous Optimization

Have some function $f : \mathbb{R}^d \to \mathbb{R}$. Want to find $\mathbf{x}^* \in R^d$ such that:

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

Or at least $\hat{\mathbf{x}} \in \mathbb{R}^d$ which is close to a minimum. E.g.

$$f(\hat{\mathbf{x}}) \leq \min_{\mathbf{x}} f(\mathbf{x}) + \epsilon.$$

Next Unit: Continuous Optimization

Have some function $f : \mathbb{R}^d \to \mathbb{R}$. Want to find $\mathbf{x}^* \in R^d$ such that:

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

Or at least $\hat{\mathbf{x}} \in \mathbb{R}^d$ which is close to a minimum. E.g.

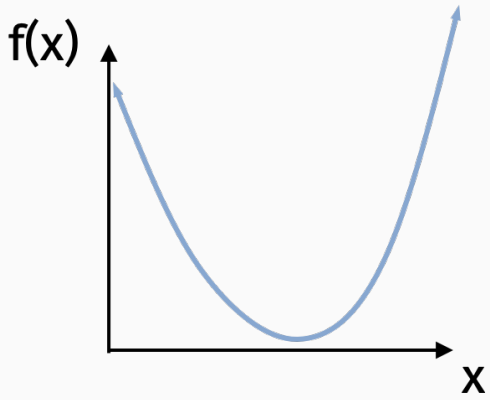
$$f(\hat{\mathbf{x}}) \leq \min_{\mathbf{x}} f(\mathbf{x}) + \epsilon.$$

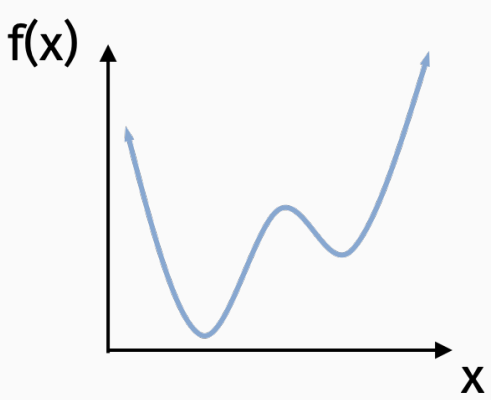
Often we have some additional constraints:

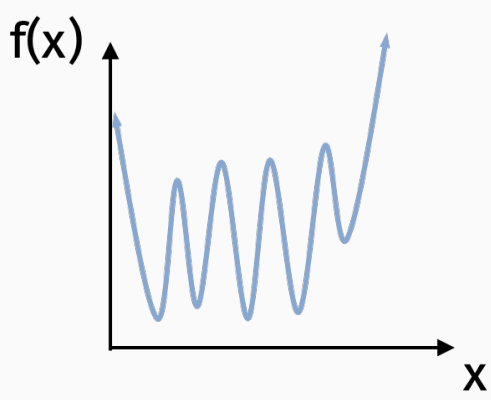
- $\mathbf{x} > 0$. (entry-wise positive)
- $\|\mathbf{x}\|_2 \leq R$, $\|\mathbf{x}\|_1 \leq R$. (convex constraints)
- $\mathbf{a}^T\mathbf{x} = c$. (linear subspace)

Continuous Optimization

Dimension $d = 1$:

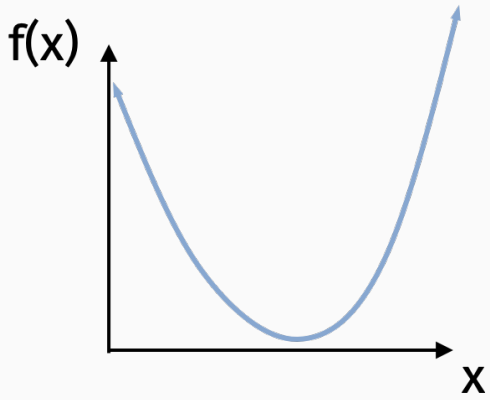


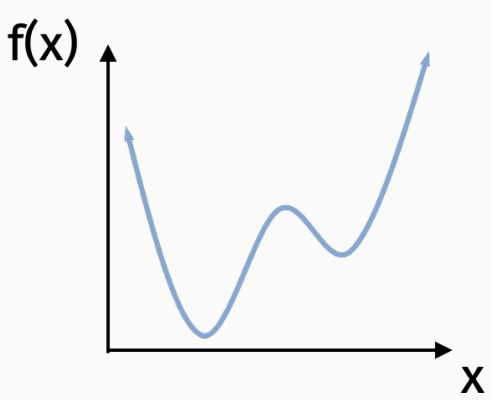


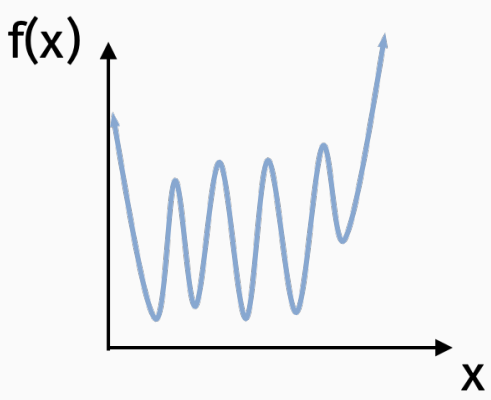


Continuous Optimization

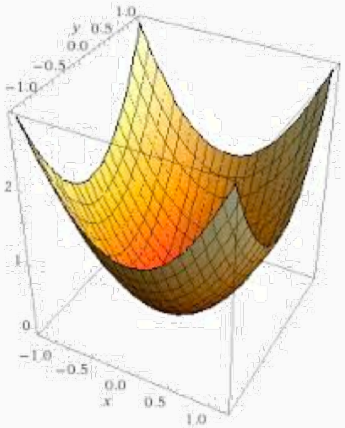
Dimension $d = 1$:

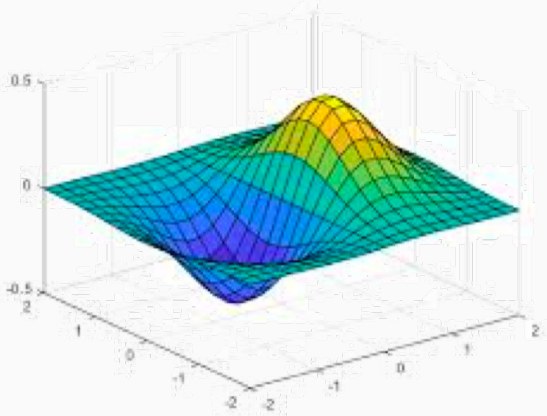


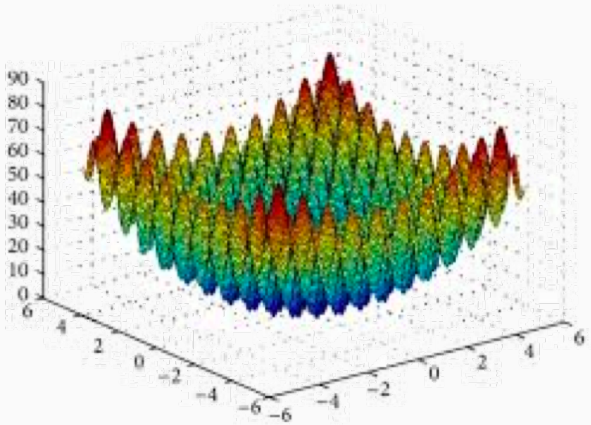




Dimension $d = 2$:







Optimization in Machine Learning

Continuous optimization is the foundation of modern machine learning.

Optimization in Machine Learning

Continuous optimization is the foundation of modern machine learning.

Supervised learning: Want to learn a model that maps inputs

- numerical data vectors
- images, video
- a sequence of tokens/words

to predictions

- numerical value (probability stock price increases)
- label (does the image contain a car? what is the next token in the sequence?)
- decision (turn car left, rotate robotic arm)

How do we map Supervised Learning back to Continuous Optimization?

Machine Learning Model

Let $M_{\mathbf{x}}$ be a model with parameters $\mathbf{x} = \{x_1, \ldots, x_k\}$, which takes as input a data vector $\mathbf{a}$ and outputs a prediction.

Machine Learning Model

Let $M_{\mathbf{x}}$ be a model with parameters $\mathbf{x} = \{x_1, \ldots, x_k\}$, which takes as input a data vector $\mathbf{a}$ and outputs a prediction.

Example:

$$M_{\mathbf{x}}(\mathbf{a}) = \text{sign}(\mathbf{a}^T\mathbf{x})$$

Machine Learning Model

Let $M_{\mathbf{x}}$ be a model with parameters $\mathbf{x} = \{x_1, \ldots, x_k\}$, which takes as input a data vector $\mathbf{a}$ and outputs a prediction.

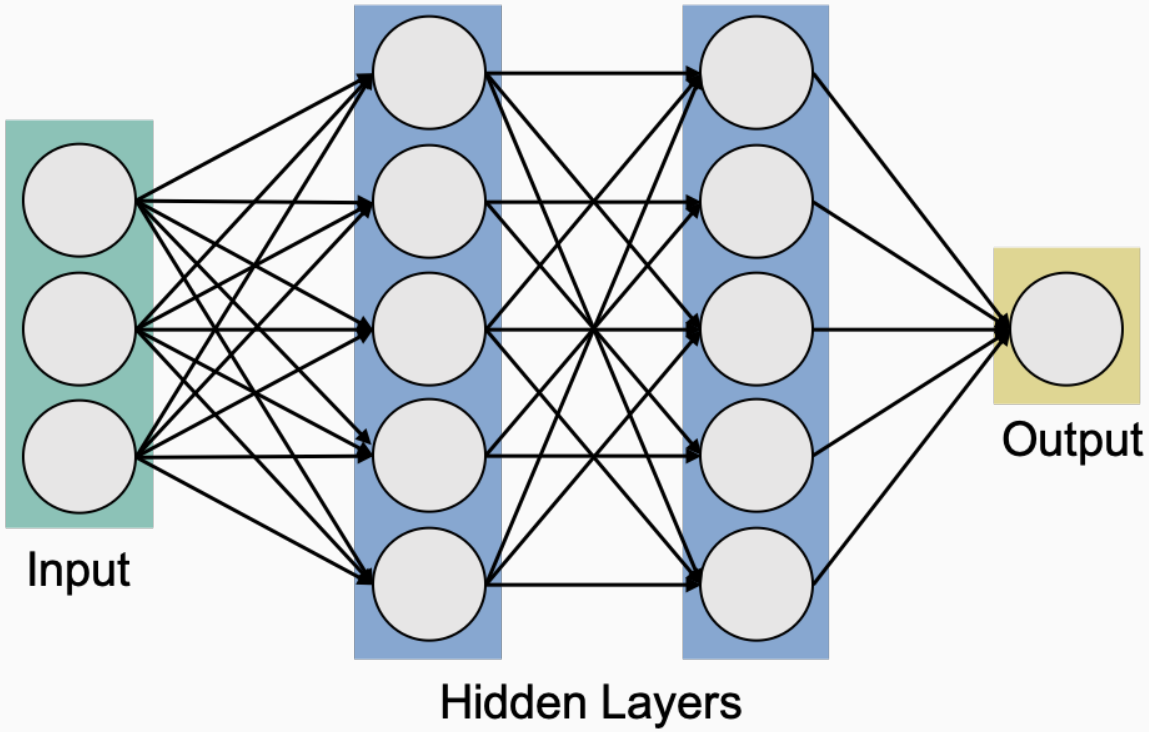
Example:

$$M_{\mathbf{x}}(\mathbf{a}) = \text{sign}(\mathbf{a}^T\mathbf{x})$$

What model is this?

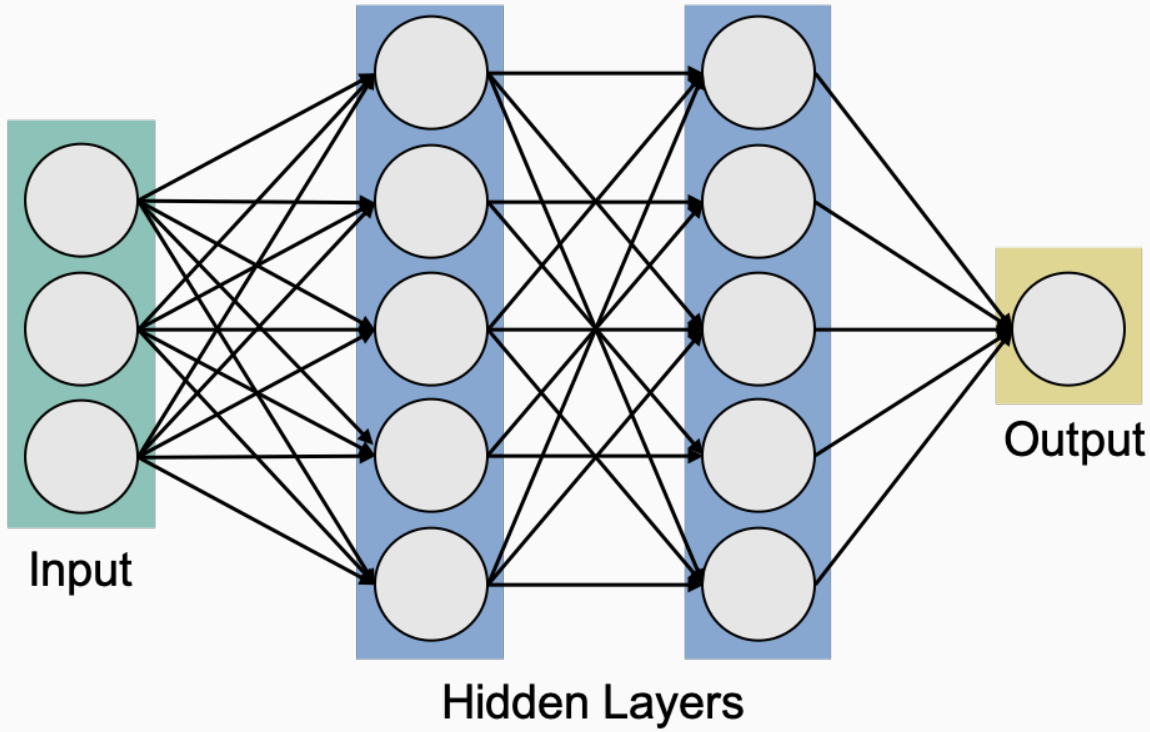
Machine Learning Model

Example: What is the parameter vector **x** here:



Machine Learning Model

Example: What is the parameter vector $\mathbf{x}$ here:



$\mathbf{x} \in \mathbb{R}^{(\#\text{ of connections})}$ is the parameter vector containing all the network weights and biases.

Supervised Learning

Classic approach in supervised learning: Find a model that works well on data that you already have the answer for (labels, values, classes, etc.).

Supervised Learning

Classic approach in supervised learning: Find a model that works well on data that you already have the answer for (labels, values, classes, etc.).

- Model $M_{\mathbf{x}}$ parameterized by a vector of numbers $\mathbf{x}$.

Supervised Learning

Classic approach in supervised learning: Find a model that works well on data that you already have the answer for (labels, values, classes, etc.).

- Model $M_{\mathbf{x}}$ parameterized by a vector of numbers $\mathbf{x}$.
- Dataset $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)}$ with outputs $y^{(1)}, \ldots, y^{(n)}$.

Want to find $\hat{\mathbf{x}}$ so that $M_{\hat{\mathbf{x}}}(\mathbf{a}^{(i)}) \approx y^{(i)}$ for $i \in 1, \ldots, n$.

(Fit the model to the training data)

Supervised Learning

Classic approach in supervised learning: Find a model that works well on data that you already have the answer for (labels, values, classes, etc.).

- Model $M_{\mathbf{x}}$ parameterized by a vector of numbers $\mathbf{x}$.
- Dataset $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)}$ with outputs $y^{(1)}, \ldots, y^{(n)}$.

Want to find $\hat{\mathbf{x}}$ so that $M_{\hat{\mathbf{x}}}(\mathbf{a}^{(i)}) \approx y^{(i)}$ for $i \in 1, \ldots, n$.

(Fit the model to the training data)

How do we turn this into a function minimization problem?

Loss Function

Loss function $L\left(M_{\mathbf{x}}(\mathbf{a}), y\right)$**:** Some measure of distance between prediction $M_{\mathbf{x}}(\mathbf{a})$ and target output y.

Loss Function

Loss function $L\left(M_{\mathbf{x}}(\mathbf{a}), y\right)$**:** Some measure of distance between prediction $M_{\mathbf{x}}(\mathbf{a})$ and target output y.

Increases if they are further apart. Decreases if they are close.

- Squared (ℓ_2) loss: $|M_{\mathbf{x}}(\mathbf{a}) - y|^2$

Loss Function

Loss function $L(M_{\mathbf{x}}(\mathbf{a}), y)$**:** Some measure of distance between prediction $M_{\mathbf{x}}(\mathbf{a})$ and target output y.

Increases if they are further apart. Decreases if they are close.

- Squared (ℓ_2) loss: $|M_{\mathbf{x}}(\mathbf{a}) - y|^2$
- Absolute deviation (ℓ_1) loss: $|M_{\mathbf{x}}(\mathbf{a}) - y|$

Loss Function

Loss function $L(M_{\mathbf{x}}(\mathbf{a}), y)$**:** Some measure of distance between prediction $M_{\mathbf{x}}(\mathbf{a})$ and target output y.

Increases if they are further apart. Decreases if they are close.

- Squared (ℓ_2) loss: $|M_{\mathbf{x}}(\mathbf{a}) - y|^2$
- Absolute deviation (ℓ_1) loss: $|M_{\mathbf{x}}(\mathbf{a}) - y|$
- Hinge loss: 1 - $y \cdot M_{\mathbf{x}}(\mathbf{a})$

Loss Function

Loss function $L\left(M_{\mathbf{x}}(\mathbf{a}), y\right)$**:** Some measure of distance between prediction $M_{\mathbf{x}}(\mathbf{a})$ and target output y.

Increases if they are further apart. Decreases if they are close.

- Squared (ℓ_2) loss: $|M_{\mathbf{x}}(\mathbf{a}) - y|^2$
- Absolute deviation (ℓ_1) loss: $|M_{\mathbf{x}}(\mathbf{a}) - y|$
- Hinge loss: 1 - $y \cdot M_{\mathbf{x}}(\mathbf{a})$
- Cross-entropy loss (log loss):
  $-o \log(M_{\mathbf{x}}(\mathbf{y})) + (1 - o) \log(1 - M_{\mathbf{x}}(\mathbf{y}))$

Empirical Risk Minimization

Input: A training dataset $(\mathbf{a}^{(1)}, y^{(1)}) \ldots, (\mathbf{a}^{(n)}, y^{(n)})$.

Empirical Risk Minimization

Input: A training dataset $(\mathbf{a}^{(1)}, y^{(1)}) \ldots, (\mathbf{a}^{(n)}, y^{(n)})$.

Loss Function:

$$f(\mathbf{x}) = \sum_{i=1}^{n} L\left(M_{\mathbf{x}}(\mathbf{a}^{(i)}), y^{(i)}\right)$$

Empirical Risk Minimization

Input: A training dataset $(\mathbf{a}^{(1)}, y^{(1)}) \ldots, (\mathbf{a}^{(n)}, y^{(n)})$.

Loss Function:

$$f(\mathbf{x}) = \sum_{i=1}^{n} L\left(M_{\mathbf{x}}(\mathbf{a}^{(i)}), y^{(i)}\right)$$

Solve the optimization problem $\min_{\mathbf{x}} f(\mathbf{x})$.

Example: Least Squares Regression

- $M_{\mathbf{x}}(\mathbf{a}) = \mathbf{x}^T\mathbf{a}$. ($\mathbf{x}$ contains the regression coefficients.)

Example: Least Squares Regression

- $M_{\mathbf{x}}(\mathbf{a}) = \mathbf{x}^T\mathbf{a}$. ($\mathbf{x}$ contains the regression coefficients.)
- $L(z, y) = |z - y|^2$.

Example: Least Squares Regression

- $M_{\mathbf{x}}(\mathbf{a}) = \mathbf{x}^T\mathbf{a}$. ($\mathbf{x}$ contains the regression coefficients.)
- $L(z, y) = |z - y|^2$.
- $f(\mathbf{x}) = \sum_{i=1}^{n} |\mathbf{x}^T\mathbf{a}^{(i)} - y^{(i)}|^2$

Example: Least Squares Regression

- $M_{\mathbf{x}}(\mathbf{a}) = \mathbf{x}^T\mathbf{a}$. ($\mathbf{x}$ contains the regression coefficients.)
- $L(z, y) = |z - y|^2$.
- $f(\mathbf{x}) = \sum_{i=1}^{n} |\mathbf{x}^T\mathbf{a}^{(i)} - y^{(i)}|^2$

Equivalent Reformulation:

$$f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$$

where $\mathbf{A}$ is a matrix with $\mathbf{a}^{(i)}$ as its i^{th} row and $\mathbf{y}$ is a vector with $y^{(i)}$ as its i^{th} entry.

Algorithms for Continuous Optimization

The choice of algorithm to minimize $f(\mathbf{x})$ will depend on:

- The form of $f(\mathbf{x})$ (is it linear, is it quadratic, does it have finite sum structure, etc.)

Algorithms for Continuous Optimization

The choice of algorithm to minimize $f(\mathbf{x})$ will depend on:

- The form of $f(\mathbf{x})$ (is it linear, is it quadratic, does it have finite sum structure, etc.)
- If there are any additional constraints imposed on $\mathbf{x}$. E.g. $\|\mathbf{x}\|_2 \leq c$.

Algorithms for Continuous Optimization

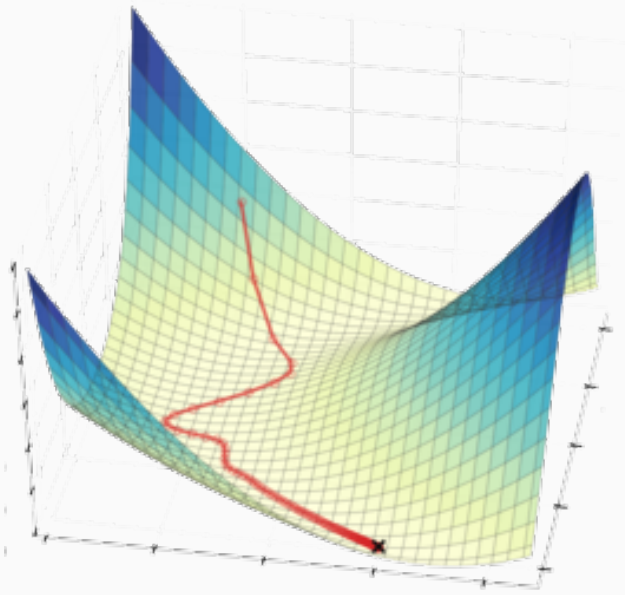
The choice of algorithm to minimize $f(\mathbf{x})$ will depend on:

- The form of $f(\mathbf{x})$ (is it linear, is it quadratic, does it have finite sum structure, etc.)
- If there are any additional constraints imposed on $\mathbf{x}$. E.g. $\|\mathbf{x}\|_2 \leq c$.

What are some example algorithms for continuous optimization?

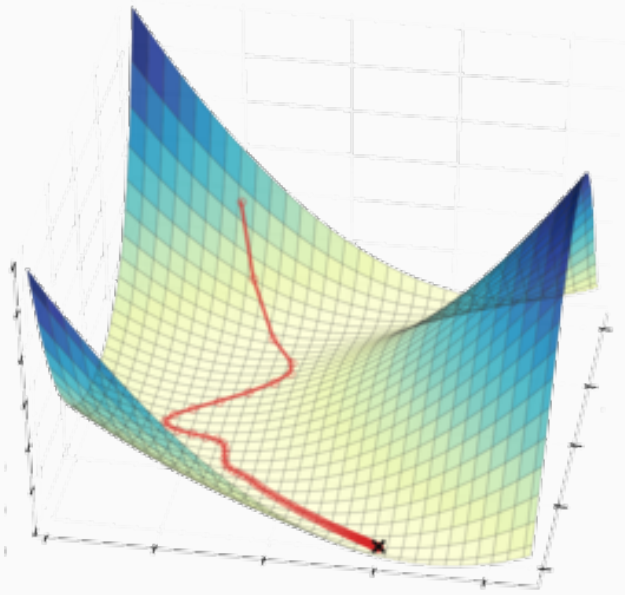
First Topic: Gradient Descent + Variants

Gradient descent: A greedy algorithm for minimizing functions of multiple variables that often works amazingly well.



First Topic: Gradient Descent + Variants

Gradient descent: A greedy algorithm for minimizing functions of multiple variables that often works amazingly well.



Runtime generally scales linearly with the dimension of $\mathbf{x}$ (although this is a bit of an over-simplification).

Second Topic: Methods Suitable for Lower Dimensions

- Cutting plane methods (e.g. center-of-gravity, ellipsoid)
- Interior point methods

Second Topic: Methods Suitable for Lower Dimensions

- Cutting plane methods (e.g. center-of-gravity, ellipsoid)
- Interior point methods

Faster and more accurate in low-dimensions, slower in very high dimensions. Generally runtime scales polynomially with the dimension of $\mathbf{x}$ (e.g., $O(d^3)$).

Calculus Review

For $i = 1, \ldots, d$, let x_i be the i^{th} entry of $\mathbf{x}$. Let $\mathbf{e}^{(i)}$ be the i^{th} standard basis vector.

Partial derivative:

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{e}^{(i)}) - f(\mathbf{x})}{t}$$

Calculus Review

For $i = 1, \ldots, d$, let x_i be the i^{th} entry of $\mathbf{x}$. Let $\mathbf{e}^{(i)}$ be the i^{th} standard basis vector.

Partial derivative:

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{e}^{(i)}) - f(\mathbf{x})}{t}$$

Directional derivative:

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t}$$

Calculus Review

For $i = 1, \ldots, d$, let x_i be the i^{th} entry of $\mathbf{x}$. Let $\mathbf{e}^{(i)}$ be the i^{th} standard basis vector.

Partial derivative:

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{e}^{(i)}) - f(\mathbf{x})}{t}$$

Directional derivative:

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t}$$

The partial derivative for coordinate 1 is the directional derivative along what direction?

Calculus Review

Gradient:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\mathbf{x}) \end{bmatrix}$$

Calculus Review

Gradient:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\mathbf{x}) \end{bmatrix}$$

Directional derivative:

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} = \nabla f(\mathbf{x})^T \mathbf{v}.$$

First Order Optimization

Helpful to think about the first order oracle model for optimization

Given a function f to minimize, assume we have:

- **Function oracle**: Evaluate $f(\mathbf{x})$ for any $\mathbf{x}$.

First Order Optimization

Helpful to think about the first order oracle model for optimization

Given a function f to minimize, assume we have:

- **Function oracle**: Evaluate $f(\mathbf{x})$ for any $\mathbf{x}$.
- **Gradient oracle**: Evaluate $\nabla f(\mathbf{x})$ for any $\mathbf{x}$.

First Order Optimization

Helpful to think about the first order oracle model for optimization

Given a function f to minimize, assume we have:

- **Function oracle**: Evaluate $f(\mathbf{x})$ for any $\mathbf{x}$.
- **Gradient oracle**: Evaluate $\nabla f(\mathbf{x})$ for any $\mathbf{x}$.

We view the implementation of these oracles as black-boxes, but they can often require a fair bit of computation.

Example Gradient Evaluation

Linear least-squares regression:

- Given $\mathbf{a}^{(1)}, \ldots \mathbf{a}^{(n)} \in \mathbb{R}^d$, $y^{(1)}, \ldots y^{(n)} \in \mathbb{R}$.
- Want to minimize:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2.$$

Example Gradient Evaluation

Linear least-squares regression:

- Given $\mathbf{a}^{(1)}, \ldots \mathbf{a}^{(n)} \in \mathbb{R}^d$, $y^{(1)}, \ldots y^{(n)} \in \mathbb{R}$.
- Want to minimize:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2.$$

What is the time complexity to implement a function oracle for $f(\mathbf{x})$?

Example Gradient Evaluation

Linear least-squares regression:

- Given $\mathbf{a}^{(1)}, \ldots \mathbf{a}^{(n)} \in \mathbb{R}^d$, $y^{(1)}, \ldots y^{(n)} \in \mathbb{R}$.
- Want to minimize:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2.$$

What is the time complexity to implement a function oracle for $f(\mathbf{x})$?

$\mathcal{O}(nd)$ (linear in d)

Example Gradient Evaluation

Linear least-squares regression:

- Want to minimize:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2.$$

Partial Derivate:

$$\frac{\partial f}{\partial x_j} = \frac{\partial}{\partial x_j} \left(\sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 \right)$$

Example Gradient Evaluation

Linear least-squares regression:

- Want to minimize:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2.$$

Partial Derivate:

$$\begin{aligned} \frac{\partial f}{\partial x_j} &= \frac{\partial}{\partial x_j} \left(\sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 \right) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial x_j} \left(\left(\mathbf{x}_1 a_1^{(i)} + \mathbf{x}_2 a_2^{(i)} + \ldots + \mathbf{x}_d a_d^{(i)} - y^{(i)} \right)^2 \right) \end{aligned}$$

Example Gradient Evaluation

Linear least-squares regression:

- Want to minimize:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2.$$

Partial Derivate:

$$\begin{aligned}
\frac{\partial f}{\partial x_j} &= \frac{\partial}{\partial x_j} \left(\sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 \right) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial x_j} \left(\left(\mathbf{x}_1 a_1^{(i)} + \mathbf{x}_2 a_2^{(i)} + \ldots + \mathbf{x}_d a_d^{(i)} - y^{(i)} \right)^2 \right) \\
&= \sum_{i=1}^{n} 2 \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right) \cdot a_j^{(i)} = 2 \boldsymbol{\alpha}^{(j)^T} (\mathbf{A}\mathbf{x} - \mathbf{y})
\end{aligned}$$

Example Gradient Evaluation

Linear least-squares regression:

- Want to minimize:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2.$$

Partial Derivate:

$$\begin{aligned}
\frac{\partial f}{\partial x_j} &= \frac{\partial}{\partial x_j} \left(\sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 \right) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial x_j} \left(\left(\mathbf{x}_1 a_1^{(i)} + \mathbf{x}_2 a_2^{(i)} + \ldots + \mathbf{x}_d a_d^{(i)} - y^{(i)} \right)^2 \right) \\
&= \sum_{i=1}^{n} 2 \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right) \cdot a_j^{(i)} = 2 \boldsymbol{\alpha}^{(j)^T} (\mathbf{A}\mathbf{x} - \mathbf{y})
\end{aligned}$$

Example Gradient Evaluation

Linear least-squares regression:

- Want to minimize:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2.$$

Partial Derivate:

$$\begin{aligned}
\frac{\partial f}{\partial x_j} &= \frac{\partial}{\partial x_j} \left(\sum_{i=1}^{n} \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right)^2 \right) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial x_j} \left(\left(\mathbf{x}_1 a_1^{(i)} + \mathbf{x}_2 a_2^{(i)} + \ldots + \mathbf{x}_d a_d^{(i)} - y^{(i)} \right)^2 \right) \\
&= \sum_{i=1}^{n} 2 \left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)} \right) \cdot a_j^{(i)} = 2 \boldsymbol{\alpha}^{(j)^T} (\mathbf{A}\mathbf{x} - \mathbf{y})
\end{aligned}$$

where $\boldsymbol{\alpha}^{(j)}$ is the j^{th} column of $\mathbf{A}$.

Example Gradient Evaluation

Linear least-squares regression:

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^{n} 2\left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)}\right) \cdot a_j^{(i)} = 2\boldsymbol{\alpha}^{(j)\,T}(\mathbf{A}\mathbf{x} - \mathbf{y})$$

where $\boldsymbol{\alpha}^{(j)}$ is the j^{th} column of $\mathbf{A}$.

$$\nabla f(\mathbf{x}) = 2\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{y})$$

What is the time complexity of a gradient oracle for $\nabla f(\mathbf{x})$?

Example Gradient Evaluation

Linear least-squares regression:

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^{n} 2\left(\mathbf{x}^T \mathbf{a}^{(i)} - y^{(i)}\right) \cdot a_j^{(i)} = 2\boldsymbol{\alpha}^{(j)\,T}(\mathbf{A}\mathbf{x} - \mathbf{y})$$

where $\boldsymbol{\alpha}^{(j)}$ is the j^{th} column of $\mathbf{A}$.

$$\nabla f(\mathbf{x}) = 2\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{y})$$

What is the time complexity of a gradient oracle for $\nabla f(\mathbf{x})$?

$$\mathcal{O}(nd)$$

Descent Methods

Greedy approach: Given a starting point $\mathbf{x}$, make a small adjustment that decreases $f(\mathbf{x})$. In particular, $\mathbf{x} \leftarrow \mathbf{x} + \eta\mathbf{v}$.

What property do I want in $\mathbf{v}$?

Descent Methods

Greedy approach: Given a starting point $\mathbf{x}$, make a small adjustment that decreases $f(\mathbf{x})$. In particular, $\mathbf{x} \leftarrow \mathbf{x} + \eta\mathbf{v}$.

What property do I want in $\mathbf{v}$?

Leading question: When η is small, what's an approximation for $f(\mathbf{x} + \eta\mathbf{v}) - f(\mathbf{x})$?

$$f(\mathbf{x} + \eta\mathbf{v}) - f(\mathbf{x}) \approx \eta \cdot \nabla f(\mathbf{x})^T \mathbf{v}$$

Directional Derivatives

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} = \nabla f(\mathbf{x})^T \mathbf{v}.$$

So:

$$f(\mathbf{x} + \eta \mathbf{v}) - f(\mathbf{x}) \approx \eta \cdot \nabla f(\mathbf{x})^T \mathbf{v}.$$

How should we choose $\mathbf{v}$ so that $f(\mathbf{x} + \eta \mathbf{v}) < f(\mathbf{x})$?

Directional Derivatives

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} = \nabla f(\mathbf{x})^T \mathbf{v}.$$

So:

$$f(\mathbf{x} + \eta \mathbf{v}) - f(\mathbf{x}) \approx \eta \cdot \nabla f(\mathbf{x})^T \mathbf{v}.$$

How should we choose v so that $f(\mathbf{x} + \eta \mathbf{v}) < f(\mathbf{x})$?

One choice: pick $v = -\nabla f(\mathbf{x})$. Then,

$$\eta \cdot \nabla f(\mathbf{x})^T \mathbf{v} = -\eta \cdot \nabla f(\mathbf{x})^T f(\mathbf{x}) = -\eta \cdot \|\nabla f(\mathbf{x})\|_2^2$$

Gradient Descent

Prototype algorithm:

- Choose starting point $\mathbf{x}^{(0)}$.

Gradient Descent

Prototype algorithm:

- Choose starting point $\mathbf{x}^{(0)}$.
- For $i = 0, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$

Gradient Descent

Prototype algorithm:

- Choose starting point $\mathbf{x}^{(0)}$.
- For $i = 0, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\mathbf{x}^{(T)}$.

Gradient Descent

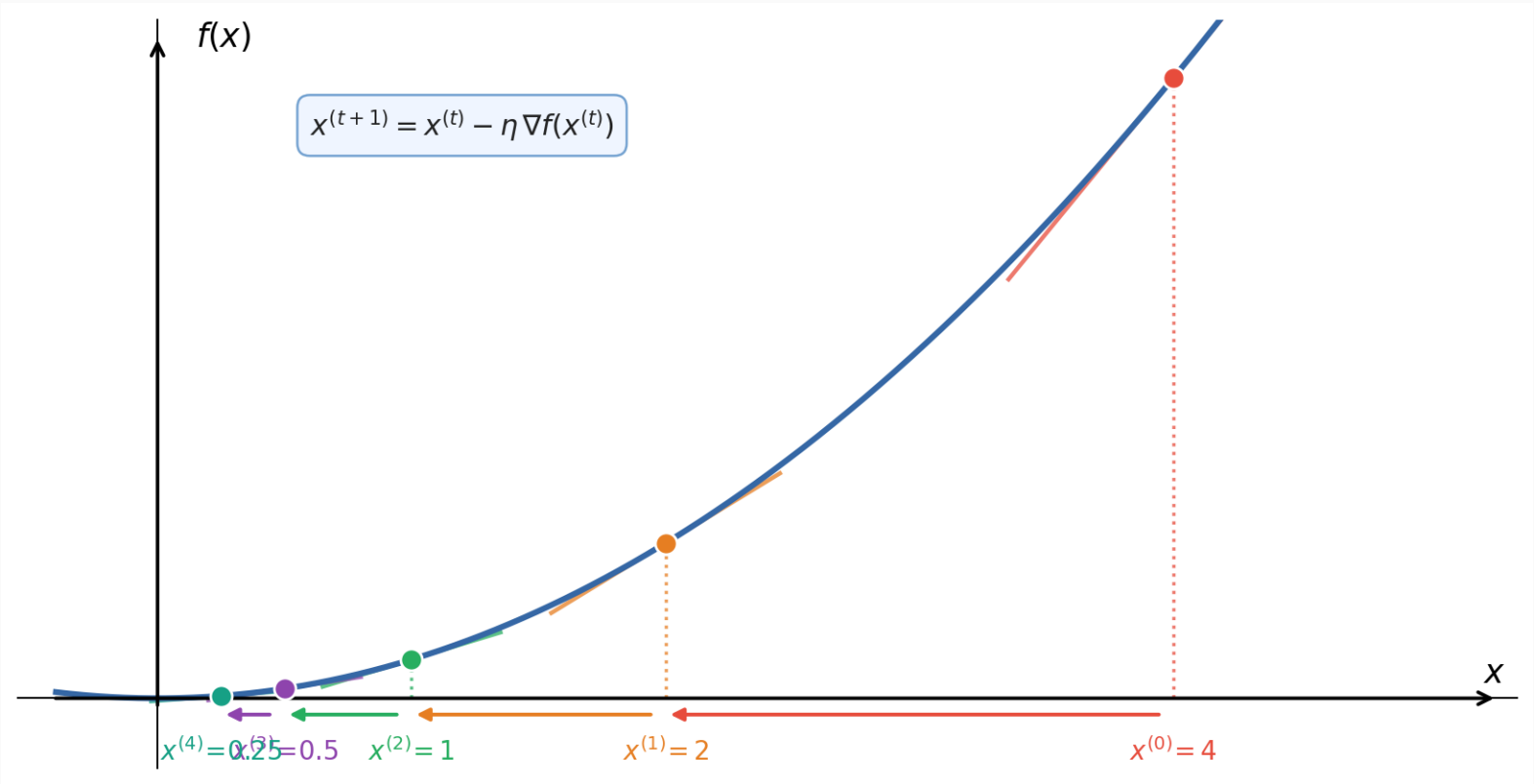
Prototype algorithm:

- Choose starting point $\mathbf{x}^{(0)}$.
- For $i = 0, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\mathbf{x}^{(T)}$.

η is a step-size parameter, which is often adapted on the go. For now, assume it is fixed ahead of time.

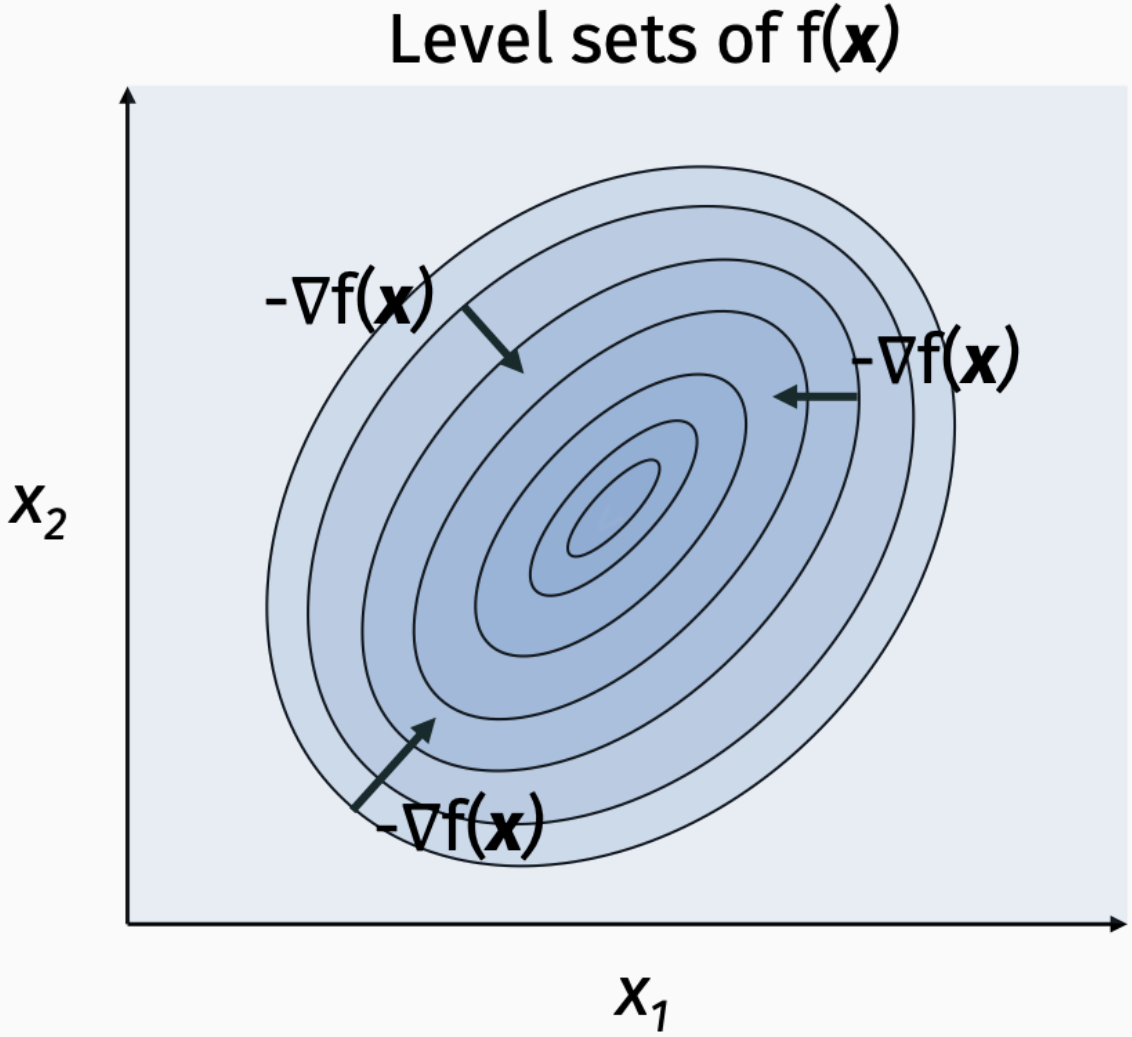
Gradient Descent Intuition

1 dimensional example: $f(x) = x^2$, starting at $x^{(0)} = 4$, step size $\eta = 0.25$.



Gradient Descent Intuition

2 dimensional example:



Key Results

For a convex function $f(\mathbf{x})$: For sufficiently small η and a sufficiently large number of iterations T, gradient descent will converge to a **near global minimum**:

$$f(\mathbf{x}^{(T)}) \leq f(\mathbf{x}^*) + \epsilon.$$

Examples: least squares regression, logistic regression, kernel regression, SVMs.

Key Results

For a convex function $f(\mathbf{x})$: For sufficiently small η and a sufficiently large number of iterations T, gradient descent will converge to a **near global minimum**:

$$f(\mathbf{x}^{(T)}) \leq f(\mathbf{x}^*) + \epsilon.$$

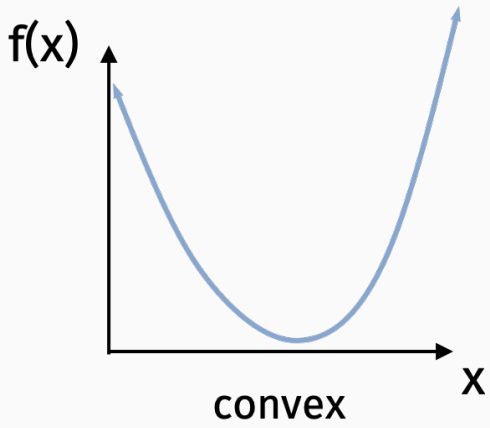
Examples: least squares regression, logistic regression, kernel regression, SVMs.

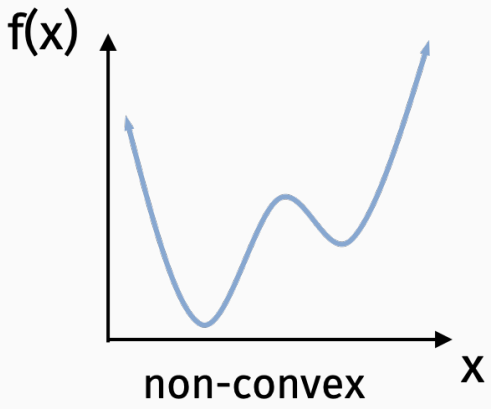
For a non-convex function $f(\mathbf{x})$: For sufficiently small η and a sufficiently large number of iterations T, gradient descent will converge to a **near stationary point** (stationary point $\nabla f(\mathbf{x}^{(T)}) = 0$):

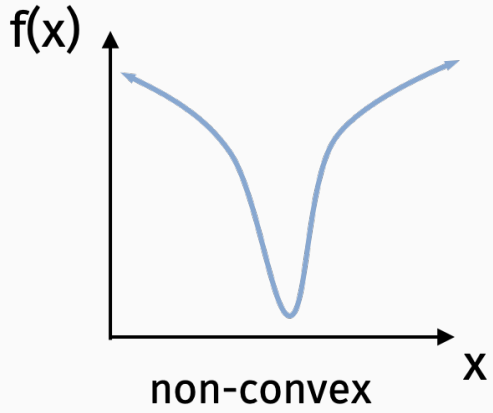
$$\|\nabla f(\mathbf{x}^{(T)})\|_2 \leq \epsilon.$$

Examples: neural networks, matrix completion problems, mixture models.

Convex vs. Non-Convex







One issue with non-convex functions is that they can have **local minima**. Even when they don't, convergence analysis requires different assumptions than convex functions.

Approach for This Unit

We care about how fast gradient descent and related methods converge, not just that they do converge.

Approach for This Unit

We care about how fast gradient descent and related methods converge, not just that they do converge.

- Bounding iteration complexity requires placing some assumptions on $f(\mathbf{x})$.

Approach for This Unit

We care about how fast gradient descent and related methods converge, not just that they do converge.

- Bounding iteration complexity requires placing some assumptions on $f(\mathbf{x})$.
- Stronger assumptions lead to better bounds on the convergence.

Understanding these assumptions can help us design faster variants of gradient descent (there are many!).

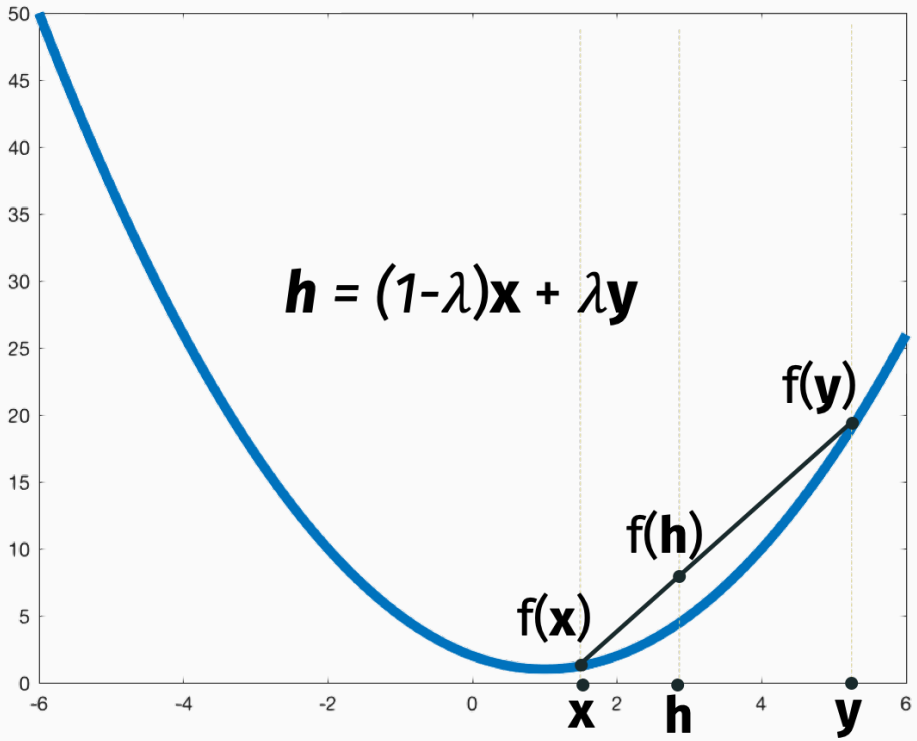
Today, we will start with **convex** functions.

Convexity

Definition (Convex)

A function f is convex iff for any $\mathbf{x}, \mathbf{y}, \lambda \in [0, 1]$:

$$(1 - \lambda) \cdot f(\mathbf{x}) + \lambda \cdot f(\mathbf{y}) \geq f\left((1 - \lambda) \cdot \mathbf{x} + \lambda \cdot \mathbf{y}\right)$$



Gradient Descent

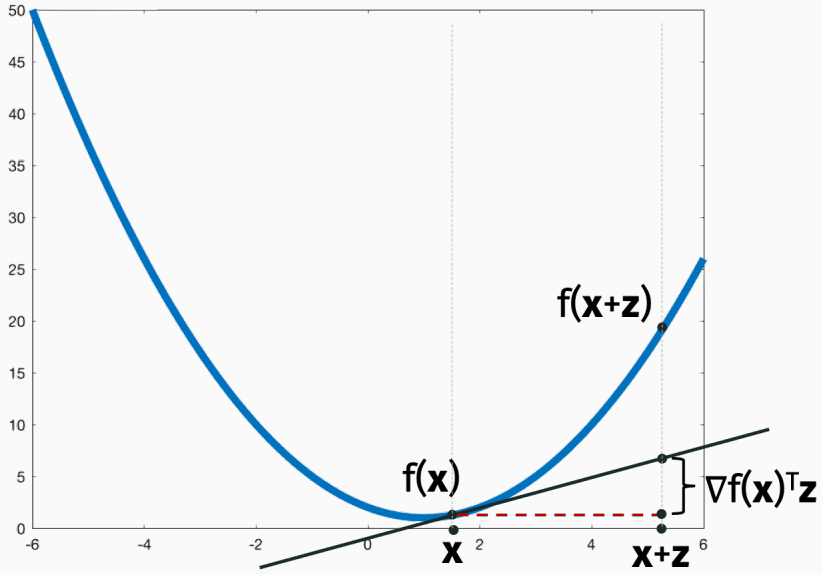
Definition (Convex)

A function f is convex if and only if for any $\mathbf{x}, \mathbf{z}$:

$$f(\mathbf{x}+\mathbf{z}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{z}$$

Equivalently: set $\mathbf{z} = y - x$, and therefore

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \nabla f(\mathbf{x})^T (\mathbf{x} - \mathbf{y})$$



Definitions of Convexity

It is easy but not obvious how to prove the equivalence between chord and tangent definitions. A short proof can be found in Karthik Sridharan's lecture notes here:

http://www.cs.cornell.edu/courses/cs6783/2018fa/lec16-supplement.pdf

Gradient Descent Analysis

Assume:

- f is convex.

Gradient Descent Analysis

Assume:

- f is convex.
- Lipschitz function: for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$.

Gradient Descent Analysis

Assume:

- f is convex.
- Lipschitz function: for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_2 \leq R$.

Gradient descent:

- Choose number of steps T.

Gradient Descent Analysis

Assume:

- f is convex.
- Lipschitz function: for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_2 \leq R$.

Gradient descent:

- Choose number of steps T.
- Starting point $\mathbf{x}^{(0)}$. E.g. $\mathbf{x}^{(0)} = \vec{0}$.

Gradient Descent Analysis

Assume:

- f is convex.
- Lipschitz function: for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_2 \leq R$.

Gradient descent:

- Choose number of steps T.
- Starting point $\mathbf{x}^{(0)}$. E.g. $\mathbf{x}^{(0)} = \vec{0}$.
- $\eta = \frac{R}{G\sqrt{T}}$

Gradient Descent Analysis

Assume:

- f is convex.
- Lipschitz function: for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_2 \leq R$.

Gradient descent:

- Choose number of steps T.
- Starting point $\mathbf{x}^{(0)}$. E.g. $\mathbf{x}^{(0)} = \vec{0}$.
- $\eta = \frac{R}{G\sqrt{T}}$
- For $i = 0, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$

Gradient Descent Analysis

Assume:

- f is convex.
- Lipschitz function: for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_2 \leq R$.

Gradient descent:

- Choose number of steps T.
- Starting point $\mathbf{x}^{(0)}$. E.g. $\mathbf{x}^{(0)} = \vec{0}$.
- $\eta = \frac{R}{G\sqrt{T}}$
- For $i = 0, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}^{(i)}} f(\mathbf{x}^{(i)})$. (return the minimum you found along the way)

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ *iterations with step-size* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ *iterations with step-size* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Proof is made tricky by the fact that $f(\mathbf{x}^{(i)})$ does not improve monotonically. We can "overshoot" the minimum.

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ *iterations with step-size* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Proof is made tricky by the fact that $f(\mathbf{x}^{(i)})$ does not improve monotonically. We can "overshoot" the minimum.

We will prove that the average solution value is low after $T = \frac{R^2 G^2}{\epsilon^2}$ iterations. I.e. that:

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ *iterations with step-size* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Proof is made tricky by the fact that $f(\mathbf{x}^{(i)})$ does not improve monotonically. We can "overshoot" the minimum.

We will prove that the average solution value is low after $T = \frac{R^2 G^2}{\epsilon^2}$ iterations. I.e. that:

$$\frac{1}{T} \sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \epsilon$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ *iterations with step-size* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Proof is made tricky by the fact that $f(\mathbf{x}^{(i)})$ does not improve monotonically. We can "overshoot" the minimum.

We will prove that the average solution value is low after $T = \frac{R^2 G^2}{\epsilon^2}$ iterations. I.e. that:

$$\frac{1}{T} \sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \epsilon$$

Of course the best solution found, $\hat{\mathbf{x}}$ is only better than the average.

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ *iterations with step-size* $\eta = \frac{R}{G\sqrt{T}}$*, then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$*.*

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ iterations with step-size $\eta = \frac{R}{G\sqrt{T}}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Interpretation: If $f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)$ is large, then in the next step we move a lot.

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ iterations with step-size $\eta = \frac{R}{G\sqrt{T}}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Interpretation: If $f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)$ is large, then in the next step we move a lot.

Claim 1(a): For all $i = 0, \ldots, T$,

$$\nabla f(\mathbf{x}^{(i)})^T (\mathbf{x}^{(i)} - \mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ iterations with step-size $\eta = \frac{R}{G\sqrt{T}}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Interpretation: If $f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)$ is large, then in the next step we move a lot.

Claim 1(a): For all $i = 0, \ldots, T$,

$$\nabla f(\mathbf{x}^{(i)})^T (\mathbf{x}^{(i)} - \mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Claim 1 follows from Claim 1(a). Why?

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ iterations with step-size $\eta = \frac{R}{G\sqrt{T}}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Interpretation: If $f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)$ is large, then in the next step we move a lot.

Claim 1(a): For all $i = 0, \ldots, T$,

$$\nabla f(\mathbf{x}^{(i)})^T (\mathbf{x}^{(i)} - \mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Claim 1 follows from Claim 1(a). Why? Convexity

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ *iterations with step size* $\eta = \frac{R}{G\sqrt{T}}$*, then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ iterations with step size $\eta = \frac{R}{G\sqrt{T}}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1(a): For all $i = 0, \ldots, T$,

$$\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \geq \nabla f(\mathbf{x}^{(i)})^T (\mathbf{x}^{(i)} - \mathbf{x}^*)$$

Recall, $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \cdot \nabla f(\mathbf{x}^{(i)})$. Plugging this in, we have

$$\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} = \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i)} - \eta \cdot \nabla f(\mathbf{x}^{(i)}) - \mathbf{x}^*\|_2^2}{2\eta}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ iterations with step size $\eta = \frac{R}{G\sqrt{T}}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1(a): For all $i = 0, \ldots, T$,

$$\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \geq \nabla f(\mathbf{x}^{(i)})^T (\mathbf{x}^{(i)} - \mathbf{x}^*)$$

Recall, $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \cdot \nabla f(\mathbf{x}^{(i)})$. Plugging this in, we have

$$\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} = \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i)} - \eta \cdot \nabla f(\mathbf{x}^{(i)}) - \mathbf{x}^*\|_2^2}{2\eta}$$

$$= \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\eta \cdot \nabla f(\mathbf{x}^{(i)})\|_2^2 + 2\eta \cdot \nabla f(\mathbf{x}^{(i)})^T (\mathbf{x}^{(i)} - \mathbf{x}^*)}{2\eta}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2 G^2}{\epsilon^2}$ iterations with step size $\eta = \frac{R}{G\sqrt{T}}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1(a): For all $i = 0, \ldots, T$,

$$\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \geq \nabla f(\mathbf{x}^{(i)})^T (\mathbf{x}^{(i)} - \mathbf{x}^*)$$

Recall, $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \cdot \nabla f(\mathbf{x}^{(i)})$. Plugging this in, we have

$$\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} = \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i)} - \eta \cdot \nabla f(\mathbf{x}^{(i)}) - \mathbf{x}^*\|_2^2}{2\eta}$$

$$= \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\eta \cdot \nabla f(\mathbf{x}^{(i)})\|_2^2 + 2\eta \cdot \nabla f(\mathbf{x}^{(i)})^T (\mathbf{x}^{(i)} - \mathbf{x}^*)}{2\eta}$$

$$= \nabla f(\mathbf{x}^{(i)})^T (\mathbf{x}^{(i)} - \mathbf{x}^*) - \frac{\eta}{2} \|\nabla f(\mathbf{x}^{(i)})\|_2^2 \geq \nabla f(\mathbf{x}^{(i)})^T (\mathbf{x}^{(i)} - \mathbf{x}^*) - \frac{\eta}{2} G^2$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Telescoping sum:

$$\sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Telescoping sum:

$$\sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$
$$+ \frac{\|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(2)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Telescoping sum:

$$\begin{aligned}\sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq & \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \\ & + \frac{\|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(2)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \\ & + \frac{\|\mathbf{x}^{(2)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(3)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}\end{aligned}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Telescoping sum:

$$\begin{aligned}\sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq & \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \\ & + \frac{\|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(2)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \\ & + \frac{\|\mathbf{x}^{(2)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(3)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \\ & \vdots\end{aligned}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Telescoping sum:

$$\begin{aligned}
\sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq & \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \\
& + \frac{\|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(2)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \\
& + \frac{\|\mathbf{x}^{(2)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(3)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \\
& \vdots \\
& + \frac{\|\mathbf{x}^{(T-1)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}
\end{aligned}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Telescoping sum:

$$\sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2 T}{2}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Telescoping sum:

$$\sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2 T}{2}$$

$$\leq \frac{R^2}{2\eta} + \frac{\eta G^2 T}{2} = \frac{RG\sqrt{T}}{2} + \frac{RG\sqrt{T}}{2} = RG\sqrt{T}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Telescoping sum:

$$\sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2 T}{2}$$

$$\leq \frac{R^2}{2\eta} + \frac{\eta G^2 T}{2} = \frac{RG\sqrt{T}}{2} + \frac{RG\sqrt{T}}{2} = RG\sqrt{T}$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$,

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Telescoping sum:

$$\sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2 T}{2}$$

$$\leq \frac{R^2}{2\eta} + \frac{\eta G^2 T}{2} = \frac{RG\sqrt{T}}{2} + \frac{RG\sqrt{T}}{2} = RG\sqrt{T}$$

Dividing both sides by T,

$$\frac{1}{T} \sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \frac{RG}{\sqrt{T}} \leq \epsilon$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Final step:

$$\frac{1}{T} \sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \epsilon$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Final step:

$$\frac{1}{T} \sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \epsilon$$

$$\implies \left[\frac{1}{T} \sum_{i=0}^{T-1} f(\mathbf{x}^{(i)}) \right] - f(\mathbf{x}^*) \leq \epsilon$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Final step:

$$\frac{1}{T} \sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \epsilon$$

$$\implies \left[\frac{1}{T} \sum_{i=0}^{T-1} f(\mathbf{x}^{(i)}) \right] - f(\mathbf{x}^*) \leq \epsilon$$

Gradient Descent Analysis

Claim (GD Convergence Bound)

If $T \geq \frac{R^2 G^2}{\epsilon^2}$ *and* $\eta = \frac{R}{G\sqrt{T}}$, *then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Final step:

$$\frac{1}{T} \sum_{i=0}^{T-1} \left[f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \right] \leq \epsilon$$

$$\implies \left[\frac{1}{T} \sum_{i=0}^{T-1} f(\mathbf{x}^{(i)}) \right] - f(\mathbf{x}^*) \leq \epsilon$$

We always have that $f(\hat{\mathbf{x}}) = \min_i f(\mathbf{x}^{(i)}) \leq \frac{1}{T} \sum_{i=0}^{T-1} f(\mathbf{x}^{(i)})$, which gives the final bound:

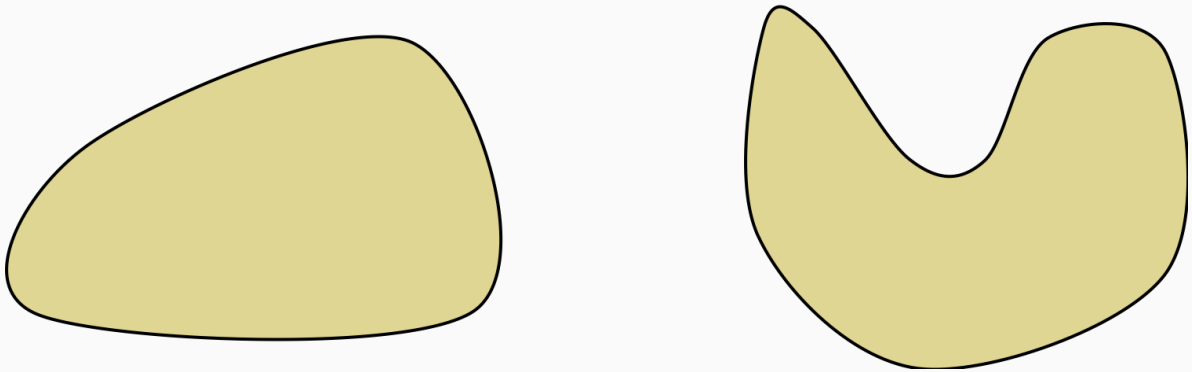
$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon.$$

Constrained Convex Optimization

Typical goal: Solve a convex minimization problem with additional convex constraints.

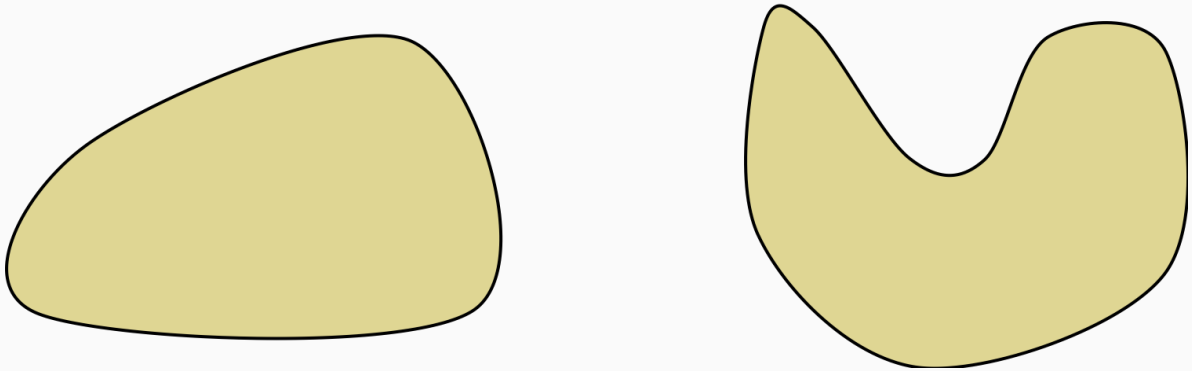
$$\min_{\mathbf{x} \in \mathcal{S}} f(\mathbf{x})$$

where $\mathcal{S}$ is a **convex set**.



Which of these is convex?

Constrained Convex Optimization



Definition (Convex set)

A set $\mathcal{S}$ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}, \lambda \in [0, 1]$:

$$(1 - \lambda)\mathbf{x} + \lambda\mathbf{y} \in \mathcal{S}.$$

Constrained Convex Optimization

Examples:

- **Norm constraint:** minimize $\|\mathbf{Ax} - \mathbf{b}\|_2$ subject to $\|\mathbf{x}\|_2 \leq \lambda$. Used e.g. for regularization, finding a sparse solution, etc.

Constrained Convex Optimization

Examples:

- **Norm constraint:** minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ subject to $\|\mathbf{x}\|_2 \leq \lambda$. Used e.g. for regularization, finding a sparse solution, etc.

- **Positivity constraint:** minimize $f(\mathbf{x})$ subject to $\mathbf{x} \geq 0$.

Constrained Convex Optimization

Examples:

- **Norm constraint:** minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ subject to $\|\mathbf{x}\|_2 \leq \lambda$. Used e.g. for regularization, finding a sparse solution, etc.

- **Positivity constraint:** minimize $f(\mathbf{x})$ subject to $\mathbf{x} \geq 0$.

- **Linear constraint:** minimize $\mathbf{c}^T\mathbf{x}$ subject to $\mathbf{A}\mathbf{x} \leq \mathbf{b}$. Linear program used in training support vector machines, industrial optimization, subroutine in integer programming, etc.

Problem with Gradient Descent

Gradient descent:

- For $i = 0, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \arg\min_i f(\mathbf{x}^{(i)})$.

Even if we start with $\mathbf{x}^{(0)} \in \mathcal{S}$, there is no guarantee that $\mathbf{x}^{(0)} - \eta \nabla f(\mathbf{x}^{(0)})$ will remain in our set.

Problem with Gradient Descent

Gradient descent:

- For $i = 0, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \arg\min_i f(\mathbf{x}^{(i)})$.

Even if we start with $\mathbf{x}^{(0)} \in \mathcal{S}$, there is no guarantee that $\mathbf{x}^{(0)} - \eta \nabla f(\mathbf{x}^{(0)})$ will remain in our set.

Extremely simple modification: Force $\mathbf{x}^{(i)}$ to be in $\mathcal{S}$ by **projecting** onto the set.

Constrained First Order Optimization

Given a function f to minimize and a convex constraint set $\mathcal{S}$, assume we have:

- **Function oracle**: Evaluate $f(\mathbf{x})$ for any $\mathbf{x}$.
- **Gradient oracle**: Evaluate $\nabla f(\mathbf{x})$ for any $\mathbf{x}$.
- **Projection oracle**: Evaluate $P_{\mathcal{S}}(\mathbf{x})$ for any $\mathbf{x}$.

$$P_{\mathcal{S}}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathcal{S}} \|\mathbf{x} - \mathbf{y}\|_2$$

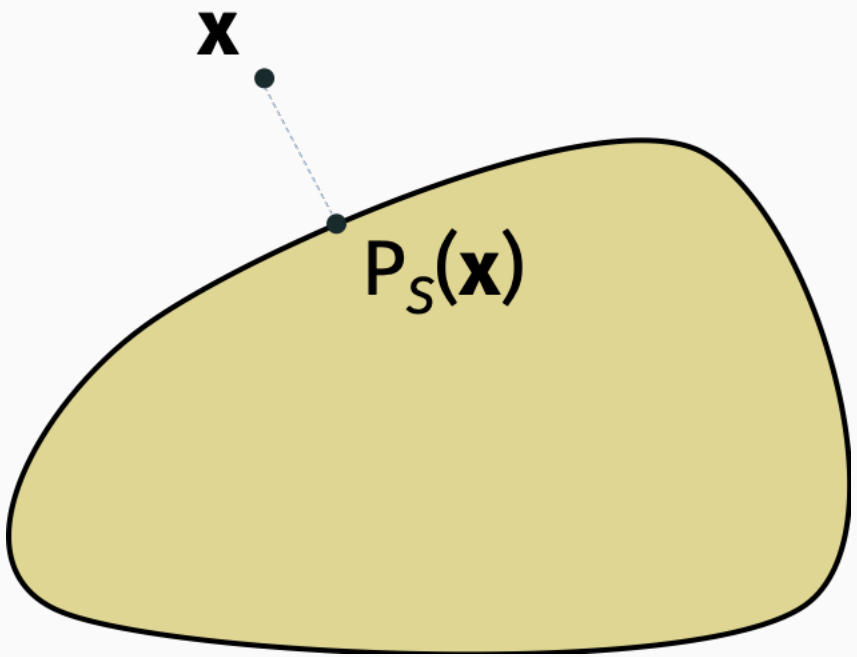
Projection Oracles

- How would you implement $P_{\mathcal{S}}$ for $\mathcal{S} = \{\mathbf{y} : \|\mathbf{y}\|_2 \leq 1\}$.

Projection Oracles

- How would you implement $P_{\mathcal{S}}$ for $\mathcal{S} = \{\mathbf{y} : \|\mathbf{y}\|_2 \leq 1\}$.
- How would you implement $P_{\mathcal{S}}$ for $\mathcal{S} = \{\mathbf{y} : \mathbf{y} = \mathbf{Q}\mathbf{z}\}$, where $Q \in \mathbb{R}^{n \times k}$

$$z^* = \arg\min_{z \in \mathbb{R}^k} \|Qz - x\|_2^2, \quad \text{thus } z^* = (Q^T Q)^{-1} Q^T x$$



Projected Gradient Descent

Given function $f(\mathbf{x})$ and set $\mathcal{S}$, such that $\|\nabla f(\mathbf{x})\|_2 \leq G$ for all $\mathbf{x} \in \mathcal{S}$ and starting point $\mathbf{x}^{(0)}$ with $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2 \leq R$.

Projected gradient descent:

- Select starting point $\mathbf{x}^{(0)}$, $\eta = \frac{R}{G\sqrt{T}}$.
- For $i = 0, \ldots, T$:
  - $\mathbf{z} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
  - $\mathbf{x}^{(i+1)} = P_{\mathcal{S}}(\mathbf{z})$
- Return $\hat{\mathbf{x}} = \arg\min_i f(\mathbf{x}^{(i)})$.

Claim (PGD Convergence Bound)

If $f, \mathcal{S}$ are convex and $T \geq \frac{R^2 G^2}{\epsilon^2}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

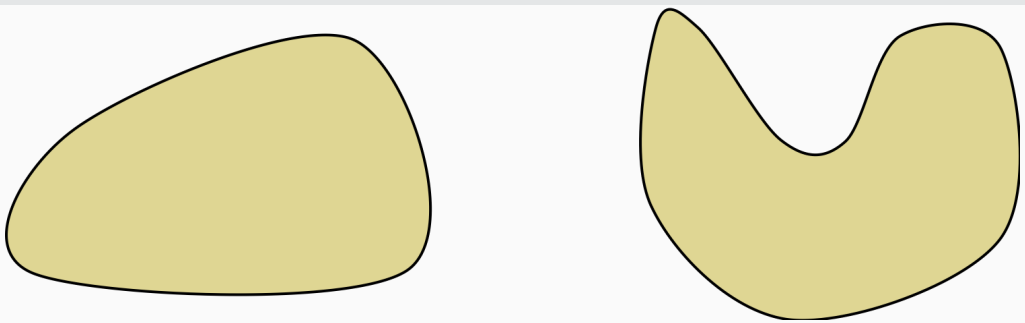
Projected Gradient Descent Analysis

Analysis is almost identical to standard gradient descent! We just need one additional claim:

Claim (Contraction Property of Convex Projection)

If $\mathcal{S}$ is convex, then for any $\mathbf{y} \in \mathcal{S}$,

$$\|\mathbf{y} - P_{\mathcal{S}}(\mathbf{x})\|_2 \leq \|\mathbf{y} - \mathbf{x}\|_2.$$



$$\|\mathbf{y} - \mathbf{x}\|_2^2 = \|\mathbf{y} - \mathbf{x} \pm P_{\mathcal{S}}(\mathbf{x})\|_2^2$$

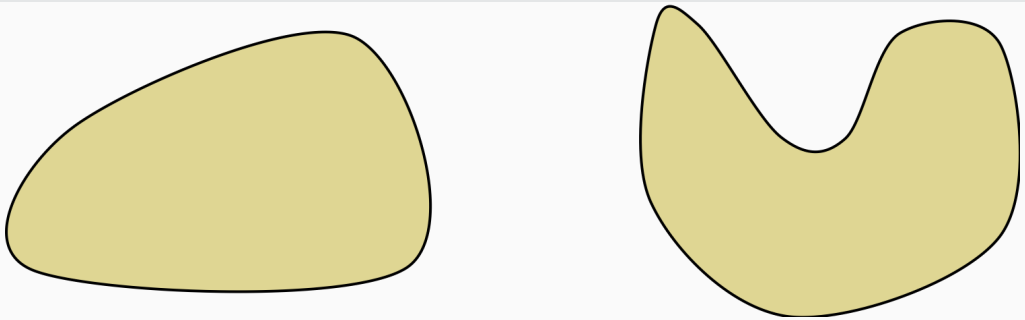
Projected Gradient Descent Analysis

Analysis is almost identical to standard gradient descent! We just need one additional claim:

Claim (Contraction Property of Convex Projection)

If $\mathcal{S}$ is convex, then for any $\mathbf{y} \in \mathcal{S}$,

$$\|\mathbf{y} - P_{\mathcal{S}}(\mathbf{x})\|_2 \leq \|\mathbf{y} - \mathbf{x}\|_2.$$



$$\begin{aligned}\|\mathbf{y} - \mathbf{x}\|_2^2 &= \|\mathbf{y} - \mathbf{x} \pm P_{\mathcal{S}}(\mathbf{x})\|_2^2 \\ &= \|\mathbf{x} - P_{\mathcal{S}}(\mathbf{x})\|_2^2 + \|\mathbf{y} - P_{\mathcal{S}}(\mathbf{x})\|_2^2 + 2\underbrace{\langle \mathbf{x} - P_{\mathcal{S}}(\mathbf{x}), P_{\mathcal{S}}(\mathbf{x}) - \mathbf{y}\rangle}_{\geq 0}\end{aligned}$$

Gradient Descent Analysis

Claim (PGD Convergence Bound)

If $f, \mathcal{S}$ are convex and $T \geq \frac{R^2 G^2}{\epsilon^2}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$, let $\mathbf{z}^{(i)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$. Then:

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{z}^{(i)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

Same telescoping sum argument:

$$\left[\frac{1}{T}\sum_{i=0}^{T-1} f(\mathbf{x}^{(i)})\right] - f(\mathbf{x}^*) \leq \frac{R^2}{2T\eta} + \frac{\eta G^2}{2}.$$

Gradient Descent Analysis

Claim (PGD Convergence Bound)

If $f, \mathcal{S}$ are convex and $T \geq \frac{R^2 G^2}{\epsilon^2}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$, let $\mathbf{z}^{(i)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$. Then:

$$\begin{aligned} f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) &\leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{z}^{(i)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \\ &\leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \end{aligned}$$

Same telescoping sum argument:

$$\left[\frac{1}{T}\sum_{i=0}^{T-1} f(\mathbf{x}^{(i)})\right] - f(\mathbf{x}^*) \leq \frac{R^2}{2T\eta} + \frac{\eta G^2}{2}.$$

Gradient Descent

Conditions:

- **Convexity:** f is a convex function, $\mathcal{S}$ is a convex set.
- **Bounded Initial Distance:**

$$\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2 \leq R$$

- **Bounded gradients (Lipschitz function)**:

$$\|\nabla f(\mathbf{x})\|_2 \leq G \text{ for all } \mathbf{x} \in \mathcal{S}.$$

Theorem (GD Convergence Bound)

(Projected) Gradient Descent outputs $\hat{\mathbf{x}}$ *with*

$$f(\hat{\mathbf{x}}) \leq \min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x}) + \epsilon\,, \qquad after \quad T = \frac{R^2 G^2}{\epsilon^2} \text{ iterations.}$$