COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2019.

Lecture 19

- Problem Set 3 on Spectral Methods due **this Friday at 8pm**.
- Can turn in without penalty until Sunday at 11:59pm.

Last Class:

- Intro to continuous optimization.
- Multivariable calculus review.
- Intro to Gradient Descent.

Last Class:

- Intro to continuous optimization.
- Multivariable calculus review.
- Intro to Gradient Descent.

This Class:

- Analysis of gradient descent for optimizing convex functions.
- Analysis of projected gradient descent for optimizing under constraints.

Gradient descent greedy motivation: At each step, make a small change to $\vec{\theta}^{(i-1)}$ to give $\vec{\theta}^{(i)}$, with minimum value of $f(\vec{\theta}^{(i)})$.

Gradient descent step: When the step size is small, this is approximate optimized by stepping in the opposite direction of the gradient:

$$\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} - \eta \cdot \vec{\nabla} f(\vec{\theta}^{(i-1)}).$$

Gradient descent greedy motivation: At each step, make a small change to $\vec{\theta}^{(i-1)}$ to give $\vec{\theta}^{(i)}$, with minimum value of $f(\vec{\theta}^{(i)})$.

Gradient descent step: When the step size is small, this is approximate optimized by stepping in the opposite direction of the gradient:

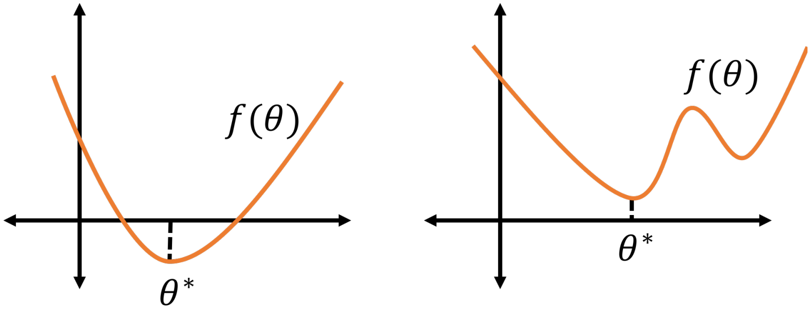
$$\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} - \eta \cdot \vec{\nabla} f(\vec{\theta}^{(i-1)}).$$

Psuedocode:

- Choose some initialization $\vec{\theta}^{(0)}$.
- For $i = 1, \ldots, t$
  - $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} - \eta \nabla f(\vec{\theta}^{(i-1)})$
- Return $\vec{\theta}^{(t)}$, as an approximate minimizer of $f(\vec{\theta})$.

Step size η is chosen ahead of time or adapted during the algorithm.

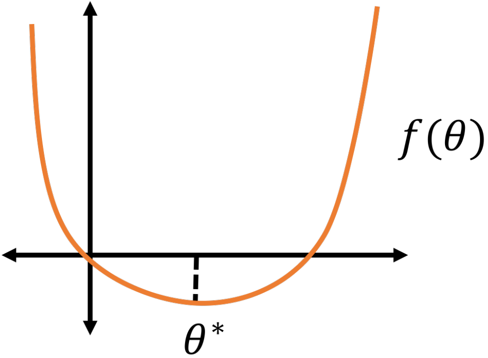
$\theta \in \mathbb{R} \quad \nabla f(\theta) \in \mathbb{R}$



Gradient Descent Update: $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} - \eta \nabla f(\vec{\theta}^{(i-1)})$

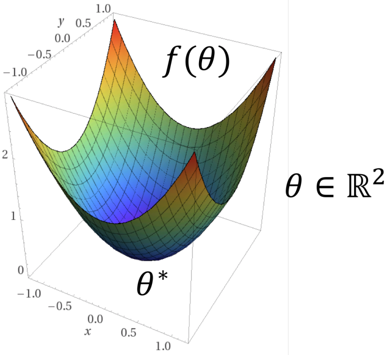
Definition – Convex Function: A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

$$(1-\lambda) \cdot f(\vec{\theta}_1) + \lambda \cdot f(\vec{\theta}_2) \geq f\left((1-\lambda) \cdot \vec{\theta}_1 + \lambda \cdot \vec{\theta}_2\right)$$



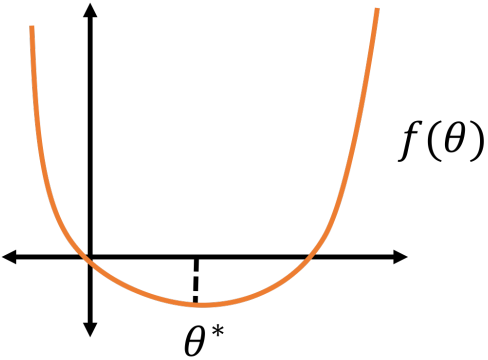
Definition – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

$$(1-\lambda)\cdot f(\vec{\theta}_1) + \lambda \cdot f(\vec{\theta}_2) \geq f\left((1-\lambda)\cdot \vec{\theta}_1 + \lambda \cdot \vec{\theta}_2\right)$$



Corollary – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

$$f(\vec{\theta}_2) - f(\vec{\theta}_1) \geq \vec{\nabla} f(\vec{\theta}_1)^T \left(\vec{\theta}_2 - \vec{\theta}_1\right)$$



We will also assume that $f(\cdot)$ is 'well-behaved' in some way.

We will also assume that $f(\cdot)$ is 'well-behaved' in some way.

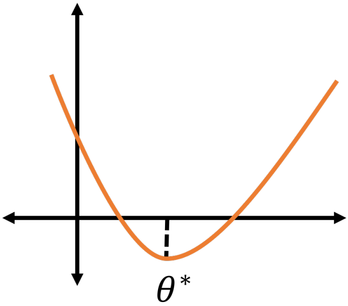
- Lipschitz (size of gradient is bounded): For all $\vec{\theta}$ and some G,

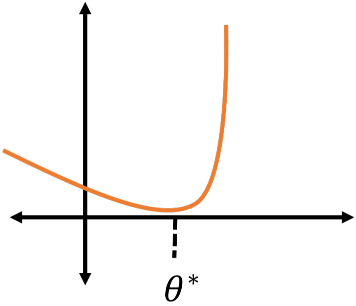
$$\|\vec{\nabla} f(\vec{\theta})\|_2 \leq G.$$

- Smooth (direction/size of gradient is not changing too quickly): For all $\vec{\theta}_1, \vec{\theta}_2$ and some β,

$$\|\vec{\nabla} f(\vec{\theta}_1) - \vec{\nabla} f(\vec{\theta}_2)\|_2 \leq \beta \cdot \|\vec{\theta}_1 - \vec{\theta}_2\|_2.$$

LIPSCHITZ ASSUMPTION





Assume that:

- f is convex.
- f is G-Lipschitz (i.e., $\|\vec{\nabla} f(\vec{\theta})\|_2 \leq G$ for all $\vec{\theta}$.)
- $\|\vec{\theta}_0 - \vec{\theta}_*\|_2 \leq R$ where θ_0 is the initialization point.

Gradient Descent

- Choose some initialization $\vec{\theta}_0$ and set $\eta = \frac{R}{G\sqrt{t}}$.
- For $i = 1, \ldots, t$
  - $\vec{\theta}_i = \vec{\theta}_{i-1} - \eta \cdot \nabla f(\vec{\theta}_{i-1})$
- Return $\hat{\theta} = \arg\min_{\vec{\theta}_0, \ldots, \vec{\theta}_t} f(\vec{\theta}_i)$.

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon.$$

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon.$$

Step 1: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$. Visually:

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon.$$

Step 1: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$. Formally:

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon.$$

Step 1: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Step 1.1: $\nabla f(\theta_i)(\theta_i - \theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon.$$

Step 1: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Step 1.1: $\nabla f(\theta_i)(\theta_i - \theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$ $\implies$ **Step 1.**

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon.$$

Step 1: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon.$$

Step 1: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$

Step 2: $\frac{1}{t}\sum_{i=1}^{t} f(\theta_i) - f(\theta_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2}$.

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon.$$

Step 2: $\frac{1}{t}\sum_{i=1}^{t} f(\theta_i) - f(\theta_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2}$.

Often want to perform convex optimization with convex constraints.

$$\theta^* = \arg\min_{\theta \in \mathcal{S}} f(\theta),$$

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

Often want to perform convex optimization with convex constraints.

$$\theta^* = \arg\min_{\theta \in \mathcal{S}} f(\theta),$$

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

Definition – Convex Set: A set $\mathcal{S} \subseteq \mathbb{R}^d$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathcal{S}$ and $\lambda \in [0, 1]$:

$$(1 - \lambda)\vec{\theta}_1 + \lambda \cdot \vec{\theta}_2 \in \mathcal{S}$$

Often want to perform convex optimization with convex constraints.

$$\theta^* = \arg\min_{\theta \in \mathcal{S}} f(\theta),$$

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

Definition – Convex Set: A set $\mathcal{S} \subseteq \mathbb{R}^d$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathcal{S}$ and $\lambda \in [0, 1]$:

$$(1 - \lambda)\vec{\theta}_1 + \lambda \cdot \vec{\theta}_2 \in \mathcal{S}$$

E.g. $\mathcal{S} = \{\vec{\theta} \in \mathbb{R}^d : \|\vec{\theta}\|_2 \leq 1\}$.

For any convex set let $P_{\mathcal{S}}(\cdot)$ denote the projection function onto $\mathcal{S}$.

- $P_{\mathcal{S}}(\vec{y}) = \arg\min_{\vec{\theta} \in \mathcal{S}} \|\vec{\theta} - \vec{y}\|_2$.

For any convex set let $P_{\mathcal{S}}(\cdot)$ denote the projection function onto $\mathcal{S}$.

- $P_{\mathcal{S}}(\vec{y}) = \arg\min_{\vec{\theta} \in \mathcal{S}} \|\vec{\theta} - \vec{y}\|_2$.
- For $\mathcal{S} = \{\vec{\theta} \in \mathbb{R}^d : \|\vec{\theta}\|_2 \leq 1\}$ what is $P_{\mathcal{S}}(\vec{y})$?

PROJECTED GRADIENT DESCENT

For any convex set let $P_{\mathcal{S}}(\cdot)$ denote the projection function onto $\mathcal{S}$.

- $P_{\mathcal{S}}(\vec{y}) = \arg\min_{\vec{\theta} \in \mathcal{S}} \|\vec{\theta} - \vec{y}\|_2$.
- For $\mathcal{S} = \{\vec{\theta} \in \mathbb{R}^d : \|\vec{\theta}\|_2 \leq 1\}$ what is $P_{\mathcal{S}}(\vec{y})$?
- For $\mathcal{S}$ being a k dimensional subspace of $\mathbb{R}^d$, what is $P_{\mathcal{S}}(\vec{y})$?

For any convex set let $P_{\mathcal{S}}(\cdot)$ denote the projection function onto $\mathcal{S}$.

- $P_{\mathcal{S}}(\vec{y}) = \arg\min_{\vec{\theta} \in \mathcal{S}} \|\vec{\theta} - \vec{y}\|_2$.
- For $\mathcal{S} = \{\vec{\theta} \in \mathbb{R}^d : \|\vec{\theta}\|_2 \leq 1\}$ what is $P_{\mathcal{S}}(\vec{y})$?
- For $\mathcal{S}$ being a k dimensional subspace of $\mathbb{R}^d$, what is $P_{\mathcal{S}}(\vec{y})$?

Projected Gradient Descent

- Choose some initialization $\vec{\theta}_0$ and set $\eta = \frac{R}{G\sqrt{t}}$.
- For $i = 1, \ldots, t$
  - $\vec{\theta}_i^{(out)} = \vec{\theta}_{i-1} - \eta \cdot \nabla f(\vec{\theta}_{i-1})$
  - $\vec{\theta}_i = P_{\mathcal{S}}(\vec{\theta}_i^{(out)})$.
- Return $\hat{\theta} = \arg\min_{\vec{\theta}_0, \ldots, \vec{\theta}_t} f(\vec{\theta}_i)$.

Visually:

Projected gradient descent can be analyzed identically to gradient descent!

Projected gradient descent can be analyzed identically to gradient descent!

Theorem – Projection to a convex set: For any convex set $\mathcal{S} \subseteq \mathbb{R}^d$, $\vec{y} \in \mathbb{R}^d$, and $\vec{\theta} \in \mathcal{S}$,

$$\|P_{\mathcal{S}}(\vec{y}) - \vec{\theta}\|_2 \leq \|\vec{y} - \vec{\theta}\|_2.$$

Theorem – Projeted GD: For convex G-Lipschitz function f, and convex set $\mathcal{S}$, Projected GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon = \min_{\theta \in \mathcal{S}} f(\theta) + \epsilon$$

Theorem – Projeted GD: For convex G-Lipschitz function f, and convex set $\mathcal{S}$, Projected GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon = \min_{\theta \in \mathcal{S}} f(\theta) + \epsilon$$

Recall: $\theta_{i+1}^{(out)} = \theta_i - \eta \cdot \nabla f(\theta_i)$ and $\theta_{i+1} = P_{\mathcal{S}}(\theta_{i+1}^{(out)})$.

Theorem – Projeted GD: For convex G-Lipschitz function f, and convex set $\mathcal{S}$, Projected GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon = \min_{\theta \in \mathcal{S}} f(\theta) + \epsilon$$

Recall: $\theta_{i+1}^{(out)} = \theta_i - \eta \cdot \nabla f(\theta_i)$ and $\theta_{i+1} = P_{\mathcal{S}}(\theta_{i+1}^{(out)})$.

Step 1: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1}^{(out)} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Theorem – Projeted GD: For convex G-Lipschitz function f, and convex set $\mathcal{S}$, Projected GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon = \min_{\theta \in \mathcal{S}} f(\theta) + \epsilon$$

Recall: $\theta_{i+1}^{(out)} = \theta_i - \eta \cdot \nabla f(\theta_i)$ and $\theta_{i+1} = P_{\mathcal{S}}(\theta_{i+1}^{(out)})$.

Step 1: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1}^{(out)} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Step 1.a: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Theorem – Projeted GD: For convex G-Lipschitz function f, and convex set $\mathcal{S}$, Projected GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of θ_*, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon = \min_{\theta \in \mathcal{S}} f(\theta) + \epsilon$$

Recall: $\theta_{i+1}^{(out)} = \theta_i - \eta \cdot \nabla f(\theta_i)$ and $\theta_{i+1} = P_{\mathcal{S}}(\theta_{i+1}^{(out)})$.

Step 1: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1}^{(out)} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Step 1.a: For all i, $f(\theta_i) - f(\theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Step 2: $\frac{1}{t}\sum_{i=1}^{t} f(\theta_i) - f(\theta_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2}$ $\implies$ Theorem.

Typical Optimization Problem in Machine Learning: Given data points $\vec{x}_1, \ldots, \vec{x}_n$ and labels/observations $y_1, \ldots, y_n$ solve:

$$\vec{\theta}_* = \arg\min_{\vec{\theta} \in \mathbb{R}^d} L(\vec{\theta}, \mathbf{X}) = \sum_{i=1}^{n} \ell(M_{\vec{\theta}}(\vec{x}_i), y_i).$$

Typical Optimization Problem in Machine Learning: Given data points $\vec{x}_1, \ldots, \vec{x}_n$ and labels/observations $y_1, \ldots, y_n$ solve:

$$\vec{\theta}_* = \arg\min_{\vec{\theta} \in \mathbb{R}^d} L(\vec{\theta}, \mathbf{X}) = \sum_{i=1}^{n} \ell(M_{\vec{\theta}}(\vec{x}_i), y_i).$$

Why is gradient descent expensive to run if you have many data points?

Typical Optimization Problem in Machine Learning: Given data points $\vec{x}_1, \ldots, \vec{x}_n$ and labels/observations $y_1, \ldots, y_n$ solve:

$$\vec{\theta}_* = \arg\min_{\vec{\theta} \in \mathbb{R}^d} L(\vec{\theta}, \mathbf{X}) = \sum_{i=1}^{n} \ell(M_{\vec{\theta}}(\vec{x}_i), y_i).$$

Why is gradient descent expensive to run if you have many data points?

$$\vec{\nabla} L(\vec{\theta}, \mathbf{X}) = \sum_{i=1}^{n} \vec{\nabla} \ell(M_{\vec{\theta}}(\vec{x}_i), y_i).$$

Typical Optimization Problem in Machine Learning: Given data points $\vec{x}_1, \ldots, \vec{x}_n$ and labels/observations $y_1, \ldots, y_n$ solve:

$$\vec{\theta}_* = \arg\min_{\vec{\theta} \in \mathbb{R}^d} L(\vec{\theta}, \mathbf{X}) = \sum_{i=1}^{n} \ell(M_{\vec{\theta}}(\vec{x}_i), y_i).$$

Why is gradient descent expensive to run if you have many data points?

$$\vec{\nabla} L(\vec{\theta}, \mathbf{X}) = \sum_{i=1}^{n} \vec{\nabla} \ell(M_{\vec{\theta}}(\vec{x}_i), y_i).$$

Solution: Take gradient step only taking into account one data point (or a small 'batch' of data points) at a time. Online and stochastic gradient descent.