COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Andrew McGregor
Lecture 23
Last Class:

- Analysis of gradient descent for optimizing convex functions.
- Introduction to convex sets and projection functions.
- (The same) analysis of projected gradient descent for optimizing under convex functions under (convex) constraints.

This Class:

- Online learning, regret, and online gradient descent.
- Application to stochastic gradient descent.
Often want to perform convex optimization with convex constraints.

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

where $S$ is a convex set.
Often want to perform convex optimization with convex constraints.

\[ \vec{\theta}^* = \arg \min_{\vec{\theta} \in S} f(\vec{\theta}), \]

where \( S \) is a convex set.

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

\[ (1 - \lambda)\vec{\theta}_1 + \lambda \cdot \vec{\theta}_2 \in S \]
Often want to perform convex optimization with convex constraints.

\[ \bar{\theta}^* = \arg\min_{\bar{\theta} \in S} f(\bar{\theta}), \]

where \( S \) is a convex set.

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

\[ (1 - \lambda)\bar{\theta}_1 + \lambda \cdot \bar{\theta}_2 \in S \]

For any convex set let \( P_S(\cdot) \) denote the projection function onto \( S \):

\[ P_S(\bar{y}) = \arg\min_{\bar{\theta} \in S} \| \bar{\theta} - \bar{y} \|_2 \]
Projected Gradient Descent

- Choose some initialization $\vec{\theta}_1$ and set $\eta = \frac{R}{G \sqrt{t}}$.
- For $i = 1, \ldots, t - 1$
  - $\vec{\theta}_{i+1}^{(out)} = \vec{\theta}_i - \eta \cdot \vec{\nabla}f(\vec{\theta}_i)$
  - $\vec{\theta}_{i+1} = P_S(\vec{\theta}_{i+1}^{(out)})$.
- Return $\hat{\theta} = \arg \min_{\vec{\theta}_i} f(\vec{\theta}_i)$. 
Analysis of projected gradient descent is almost identical to gradient descent analysis!

Theorem – Projection to a convex set:
For any convex set $S \subseteq \mathbb{R}^d$, $\vec{y} \in \mathbb{R}^d$, and $\vec{\theta} \in S$,
\[
\|P_S(\vec{y}) - \vec{\theta}\|_2 \leq \|\vec{y} - \vec{\theta}\|_2.
\]
Analysis of projected gradient descent is almost identical to gradient descent analysis!

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

$$\|P_S(\tilde{y}) - \bar{\theta}\|_2 \leq \|\tilde{y} - \bar{\theta}\|_2.$$
**Theorem – Projected GD:** For convex $G$-Lipschitz function $f$, and convex set $S$, 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 $\hat{\theta}_* = \min_{\bar{\theta} \in S} f(\bar{\theta})$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\hat{\theta}_*) + \epsilon$$
Theorem – Projected GD: For convex $G$-Lipschitz function $f$, and convex set $S$, 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 $\vec{\theta}^* = \min_{\vec{\theta} \in S} f(\vec{\theta})$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}^*) + \epsilon$$

Recall: $\vec{\theta}^{(out)}_{i+1} = \vec{\theta}_i - \eta \cdot \nabla f(\vec{\theta}_i)$ and $\vec{\theta}_{i+1} = P_S(\vec{\theta}^{(out)}_{i+1})$. 
Theorem – Projected GD: For convex $G$-Lipschitz function $f$, and convex set $S$, 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 $\bar{\theta}_* = \min_{\theta \in S} f(\theta)$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\bar{\theta}_*) + \epsilon$$

Recall: $\bar{\theta}_{i+1}^{(out)} = \bar{\theta}_i - \eta \cdot \nabla f(\bar{\theta}_i)$ and $\tilde{\theta}_{i+1} = P_S(\bar{\theta}_{i+1}^{(out)})$.

Step 1: For all $i$, $f(\bar{\theta}_i) - f(\bar{\theta}_*) \leq \frac{||\bar{\theta}_i - \theta_*||^2 - ||\bar{\theta}_{i+1}^{(out)} - \bar{\theta}_*||^2}{2\eta} + \frac{\eta G^2}{2}$. 
**Theorem – Projected GD:** For convex $G$-Lipschitz function $f$, and convex set $S$, 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 $\vec{\theta}^* = \min_{\vec{\theta} \in S} f(\vec{\theta})$, outputs $\hat{\vec{\theta}}$ satisfying:

$$f(\hat{\vec{\theta}}) \leq f(\vec{\theta}^*) + \epsilon$$

Recall: $\vec{\theta}^{(out)}_{i+1} = \vec{\theta}_i - \eta \cdot \nabla f(\vec{\theta}_i)$ and $\vec{\theta}_{i+1} = P_S(\vec{\theta}^{(out)}_{i+1})$.

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

**Step 1.a:** For all $i$, $f(\vec{\theta}_i) - f(\vec{\theta}^*) \leq \frac{||\vec{\theta}_i - \vec{\theta}^*||^2_2 - ||\vec{\theta}_{i+1} - \vec{\theta}^*||^2_2}{2\eta} + \frac{\eta G^2}{2}$.
**Theorem – Projected GD:** For convex $G$-Lipschitz function $f$, and convex set $S$, 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 $\hat{\theta}_* = \min_{\theta \in S} f(\theta)$, outputs $\hat{\theta}$ satisfying:

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

**Recall:** $\vec{\theta}^{(out)}_{i+1} = \vec{\theta}_i - \eta \cdot \nabla f(\vec{\theta}_i)$ and $\vec{\theta}_{i+1} = P_S(\vec{\theta}^{(out)}_{i+1})$.

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

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

**Step 2:** $\frac{1}{t} \sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2} \implies$ Theorem.
In reality many learning problems are online.

- Websites optimize ads or recommendations to show users, given continuous feedback from these users.
- Spam filters are incrementally updated and adapt as they see more examples of spam over time.
- Face recognition systems, other classification systems, learn from mistakes over time.
In reality many learning problems are online.

- Websites optimize ads or recommendations to show users, given continuous feedback from these users.
- Spam filters are incrementally updated and adapt as they see more examples of spam over time.
- Face recognition systems, other classification systems, learn from mistakes over time.

Want to minimize some global loss $L(\theta, X) = \sum_{i=1}^{n} \ell(\theta, \vec{x}_i)$, when data points are presented in an online fashion $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ (similar to streaming algorithms)
In reality many learning problems are online.

- Websites optimize ads or recommendations to show users, given continuous feedback from these users.
- Spam filters are incrementally updated and adapt as they see more examples of spam over time.
- Face recognition systems, other classification systems, learn from mistakes over time.

Want to minimize some global loss \( L(\theta, X) = \sum_{i=1}^{n} \ell(\theta, x_i) \), when data points are presented in an online fashion \( x_1, x_2, \ldots, x_n \) (similar to streaming algorithms)

Stochastic gradient descent is a special case: when data points are considered a random order for computational reasons.
**Online Optimization**: In place of a single function $f$, we see a different objective function at each step:

$$f_1, f_2, \ldots, f_t : \mathbb{R}^d \rightarrow \mathbb{R}$$
Online Optimization: In place of a single function $f$, we see a different objective function at each step:

$$f_1, f_2, \ldots, f_t : \mathbb{R}^d \rightarrow \mathbb{R}$$

- At each step, first pick (play) a parameter vector $\bar{\theta}^{(i)}$.
- Then are told $f_i$ and incur cost $f_i(\bar{\theta}^{(i)})$.
- **Goal:** Minimize total cost $\sum_{i=1}^{t} f_i(\bar{\theta}^{(i)})$. 
Online Optimization: In place of a single function $f$, we see a different objective function at each step:

$$f_1, f_2, \ldots, f_t : \mathbb{R}^d \rightarrow \mathbb{R}$$

- At each step, first pick (play) a parameter vector $\vec{\theta}^{(i)}$.
- Then are told $f_i$ and incur cost $f_i(\vec{\theta}^{(i)})$.
- **Goal:** Minimize total cost $\sum_{i=1}^{t} f_i(\vec{\theta}^{(i)})$.

Our analysis will make no assumptions on how $f_1, \ldots, f_t$ are related to each other!
UI design via online optimization.

- Parameter vector $\vec{\theta}(i)$: some encoding of the layout at step $i$.
- Functions $f_1, \ldots, f_t$: $f_i(\vec{\theta}(i)) = 1$ if user does not click ‘add to cart’ and $f_i(\vec{\theta}(i)) = 0$ if they do click.
- Want to maximize number of purchases, i.e., minimize $\sum_{i=1}^{t} f_i(\vec{\theta}(i))$. 
Home pricing tools.

\[\tilde{x} = [\#baths, \#beds, \#floors ...]\]

- Parameter vector \(\tilde{\theta}^{(i)}\): coefficients of linear model at step \(i\).
- Functions \(f_1, \ldots, f_t\): \(f_i(\tilde{\theta}^{(i)}) = (\langle \tilde{x}_i, \tilde{\theta}^{(i)} \rangle - \text{price}_i)^2\) revealed when \(\text{home}_i\) is listed or sold.
- Want to minimize total squared error \(\sum_{i=1}^{t} f_i(\tilde{\theta}^{(i)})\) (same as classic least squares regression).
In normal optimization, we seek $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq \min_{\tilde{\theta}} f(\tilde{\theta}) + \epsilon.$$
In normal optimization, we seek \( \hat{\theta} \) satisfying:

\[
f(\hat{\theta}) \leq \min_{\vec{\theta}} f(\vec{\theta}) + \epsilon.
\]

In online optimization we will ask for the same.

\[
\sum_{i=1}^{t} f_i(\vec{\theta}(i)) \leq \min_{\vec{\theta}} \sum_{i=1}^{t} f_i(\vec{\theta}) + \epsilon = \sum_{i=1}^{t} f_i(\vec{\theta}^{off}) + \epsilon
\]

\( \epsilon \) is called the regret and \( \epsilon/t \) is the average regret.
In normal optimization, we seek $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq \min_{\theta} f(\bar{\theta}) + \epsilon.$$ 

In online optimization we will ask for the same.

$$\sum_{i=1}^{t} f_i(\bar{\theta}^{(i)}) \leq \min_{\theta} \sum_{i=1}^{t} f_i(\bar{\theta}) + \epsilon = \sum_{i=1}^{t} f_i(\bar{\theta}^{\text{off}}) + \epsilon$$

$\epsilon$ is called the regret and $\epsilon/t$ is the average regret.

- This error metric is a bit unusual: Comparing online solution to best fixed solution in hindsight. $\epsilon$ can be negative!
What if for $i = 1, \ldots, t$, $f_i(\theta) = |\theta - 1000|$ or $f_i(\theta) = |\theta + 1000|$ in an alternating pattern?

How small can the regret $\epsilon$ be? $\sum_{i=1}^{t} f_i(\vec{\theta}(i)) \leq \sum_{i=1}^{t} f_i(\vec{\theta}^{off}) + \epsilon$. 
What if for $i = 1, \ldots, t$, $f_i(\theta) = |\theta - 1000|$ or $f_i(\theta) = |\theta + 1000|$ in an alternating pattern?

How small can the regret $\epsilon$ be? \[ \sum_{i=1}^{t} f_i(\vec{\theta}^{(i)}) \leq \sum_{i=1}^{t} f_i(\vec{\theta}^{\text{off}}) + \epsilon. \]

What if for $i = 1, \ldots, t$, $f_i(\theta) = |\theta - 1000|$ or $f_i(\theta) = |\theta + 1000|$ in no particular pattern? How can any online learning algorithm hope to achieve small regret?
Assume that:

- $f_1, \ldots, f_t$ are all convex.
- Each $f_i$ is $G$-Lipschitz (i.e., $\|\nabla f_i(\theta)\|_2 \leq G$ for all $\theta$.)
- $\|\theta^{(1)} - \bar{\theta}^{\text{off}}\|_2 \leq R$ where $\theta^{(1)}$ is the first vector chosen.
Assume that:

- $f_1, \ldots, f_t$ are all convex.
- Each $f_i$ is $G$-Lipschitz (i.e., $\|\nabla f_i(\theta)\|_2 \leq G$ for all $\theta$.)
- $\|\theta^{(1)} - \theta^{\text{off}}\|_2 \leq R$ where $\theta^{(1)}$ is the first vector chosen.

**Online Gradient Descent**

- Pick some initial $\theta^{(1)}$.
- Set step size $\eta = \frac{R}{G \sqrt{t}}$.
- For $i = 1, \ldots, t$
  - Play $\theta^{(i)}$ and incur cost $f_i(\theta^{(i)})$.
  - $\theta^{(i+1)} = \theta^{(i)} - \eta \cdot \nabla f_i(\theta^{(i)})$
Theorem – OGD on Convex Lipschitz Functions: For convex \( G \)-Lipschitz \( f_1, \ldots, f_t \), OGD initialized with starting point \( \theta^{(1)} \) within radius \( R \) of \( \theta^{\text{off}} \), using step size \( \eta = \frac{R}{G\sqrt{t}} \), has regret bounded by:

\[
\left[ \sum_{i=1}^{t} f_i(\theta^{(i)}) - \sum_{i=1}^{t} f_i(\theta^{\text{off}}) \right] \leq RG\sqrt{t}
\]
Theorem – OGD on Convex Lipschitz Functions: For convex $G$-Lipschitz $f_1, \ldots, f_t$, OGD initialized with starting point $\theta^{(1)}$ within radius $R$ of $\theta^{\text{off}}$, using step size $\eta = \frac{R}{G\sqrt{t}}$, has regret bounded by:

$$
\left[ \sum_{i=1}^{t} f_i(\theta^{(i)}) - \sum_{i=1}^{t} f_i(\theta^{\text{off}}) \right] \leq RG\sqrt{t}
$$

Upper bound on average regret goes to 0 and $t \to \infty$. 
Theorem — OGD on Convex Lipschitz Functions: For convex $G$-Lipschitz $f_1, \ldots, f_t$, OGD initialized with starting point $\theta^{(1)}$ within radius $R$ of $\theta^{\text{off}}$, using step size $\eta = \frac{R}{G\sqrt{t}}$, has regret bounded by:

$$\left[ \sum_{i=1}^{t} f_i(\theta^{(i)}) - \sum_{i=1}^{t} f_i(\theta^{\text{off}}) \right] \leq RG\sqrt{t}$$

Upper bound on average regret goes to 0 and $t \to \infty$. No assumptions on $f_1, \ldots, f_t$!
Theorem – OGD on Convex Lipschitz Functions: For convex $G$-Lipschitz $f_1, \ldots, f_t$, OGD initialized with starting point $\theta^{(1)}$ within radius $R$ of $\theta^{\text{off}}$, using step size $\eta = \frac{R}{G\sqrt{t}}$, has regret bounded by:

$$\left[ \sum_{i=1}^{t} f_i(\theta^{(i)}) - \sum_{i=1}^{t} f_i(\theta^{\text{off}}) \right] \leq RG\sqrt{t}$$

Upper bound on average regret goes to 0 and $t \to \infty$. No assumptions on $f_1, \ldots, f_t$!

Step 1.1: For all $i$, $\nabla f_i(\theta^{(i)})^T (\theta^{(i)} - \theta^{\text{off}}) \leq \frac{\|\theta^{(i)} - \theta^{\text{off}}\|_2^2 - \|\theta^{(i+1)} - \theta^{\text{off}}\|_2^2}{2\eta} + \frac{\eta G^2}{2}$
Theorem – OGD on Convex Lipschitz Functions: For convex $G$-Lipschitz $f_1, \ldots, f_t$, OGD initialized with starting point $\theta^{(1)}$ within radius $R$ of $\theta^{off}$, using step size $\eta = \frac{R}{G\sqrt{t}}$, has regret bounded by:

$$\left[ \sum_{i=1}^{t} f_i(\theta^{(i)}) - \sum_{i=1}^{t} f_i(\theta^{off}) \right] \leq RG\sqrt{t}$$

Upper bound on average regret goes to 0 and $t \to \infty$. No assumptions on $f_1, \ldots, f_t$!

**Step 1.1:** For all $i$, $\nabla f_i(\theta^{(i)})^T (\theta^{(i)} - \theta^{off}) \leq \frac{\|	heta^{(i)} - \theta^{off}\|_2^2 - \|	heta^{(i+1)} - \theta^{off}\|_2^2}{2\eta} + \frac{\eta G^2}{2}$

Convexity $\implies$ **Step 1:** For all $i$,

$$f_i(\theta^{(i)}) - f_i(\theta^{off}) \leq \frac{\|	heta^{(i)} - \theta^{off}\|_2^2 - \|	heta^{(i+1)} - \theta^{off}\|_2^2}{2\eta} + \frac{\eta G^2}{2}.$$
**Theorem – OGD on Convex Lipschitz Functions:** For convex $G$-Lipschitz $f_1, \ldots, f_t$, OGD initialized with starting point $\theta^{(1)}$ within radius $R$ of $\theta^{\text{off}}$, using step size $\eta = \frac{R}{G\sqrt{t}}$, has regret bounded by:

$$\left[ \sum_{i=1}^{t} f_i(\theta^{(i)}) - \sum_{i=1}^{t} f_i(\theta^{\text{off}}) \right] \leq RG\sqrt{t}$$

**Step 1:** For all $i$, $f_i(\theta^{(i)}) - f_i(\theta^{\text{off}}) \leq \frac{\|\theta^{(i)}-\theta^{\text{off}}\|_2^2 - \|\theta^{(i+1)}-\theta^{\text{off}}\|_2^2}{2\eta} + \frac{\eta G^2}{2}$
Theorem – OGD on Convex Lipschitz Functions: For convex $G$-Lipschitz $f_1, \ldots, f_t$, OGD initialized with starting point $\theta^{(1)}$ within radius $R$ of $\theta^{\text{off}}$, using step size $\eta = \frac{R}{G \sqrt{t}}$, has regret bounded by:

$$\left[ \sum_{i=1}^{t} f_i(\theta^{(i)}) - \sum_{i=1}^{t} f_i(\theta^{\text{off}}) \right] \leq RG\sqrt{t}$$

**Step 1:** For all $i$, $f_i(\theta^{(i)}) - f_i(\theta^{\text{off}}) \leq \frac{\|\theta^{(i)} - \theta^{\text{off}}\|_2^2 - \|\theta^{(i+1)} - \theta^{\text{off}}\|_2^2}{2\eta} + \frac{\eta G^2}{2} \implies$

$$\left[ \sum_{i=1}^{t} f_i(\theta^{(i)}) - \sum_{i=1}^{t} f_i(\theta^{\text{off}}) \right] \leq \sum_{i=1}^{t} \frac{\|\theta^{(i)} - \theta^{\text{off}}\|_2^2 - \|\theta^{(i+1)} - \theta^{\text{off}}\|_2^2}{2\eta} + \frac{t \cdot \eta G^2}{2}.$$
Stochastic gradient descent is an efficient offline optimization method, seeking $\hat{\theta}$ with

$$f(\hat{\theta}) \leq \min_{\tilde{\theta}} f(\tilde{\theta}) + \epsilon = f(\tilde{\theta}^*) + \epsilon.$$
Stochastic gradient descent is an efficient offline optimization method, seeking \( \hat{\theta} \) with

\[
f(\hat{\theta}) \leq \min_{\tilde{\theta}} f(\tilde{\theta}) + \epsilon = f(\tilde{\theta}^*) + \epsilon.
\]

- The most popular optimization method in modern machine learning.
- Easily analyzed as a special case of online gradient descent!
Assume that:

- $f$ is convex and decomposable as $f(\theta) = \sum_{j=1}^{n} f_j(\theta)$.
- E.g., $L(\theta, X) = \sum_{j=1}^{n} \ell(\theta, x_j)$. 

Stochastic Gradient Descent

1. Pick some initial $\theta(1)$.
2. Set step size $\eta = \frac{R}{G} \sqrt{t}$.
3. For $i = 1, \ldots, t$:
   1. Pick random $j_i \in \{1, \ldots, n\}$.
   2. $\theta(i+1) = \theta(i) - \eta \cdot \nabla f_{j_i}(\theta(i))$.
4. Return $\hat{\theta} = \frac{1}{t} \sum_{i=1}^{t} \theta(i)$. 


Assume that:

• $f$ is convex and decomposable as $f(\vec{\theta}) = \sum_{j=1}^{n} f_j(\vec{\theta})$.
  • E.g., $L(\vec{\theta}, X) = \sum_{j=1}^{n} \ell(\vec{\theta}, \vec{x}_j)$.
  • Each $f_j$ is $\frac{G}{n}$-Lipschitz (i.e., $\|\nabla f_j(\vec{\theta})\|_2 \leq \frac{G}{n}$ for all $\vec{\theta}$.)
  • What does this imply about how Lipschitz $f$ is?
Assume that:

• $f$ is convex and decomposable as $f(\vec{\theta}) = \sum_{j=1}^{n} f_j(\vec{\theta})$.
  - E.g., $L(\vec{\theta}, X) = \sum_{j=1}^{n} \ell(\vec{\theta}, \vec{x}_j)$.

• Each $f_j$ is $\frac{G}{n}$-Lipschitz (i.e., $\|\vec{\nabla} f_j(\vec{\theta})\|_2 \leq \frac{G}{n}$ for all $\vec{\theta}$.)
  - What does this imply about how Lipschitz $f$ is?

• Initialize with $\theta^{(1)}$ satisfying $\|\vec{\theta}^{(1)} - \vec{\theta}^*\|_2 \leq R$. 
STOCHASTIC GRADIENT DESCENT

Assume that:

• $f$ is convex and decomposable as $f(\vec{\theta}) = \sum_{j=1}^{n} f_j(\vec{\theta})$.
  • E.g., $L(\vec{\theta}, X) = \sum_{j=1}^{n} \ell(\vec{\theta}, \vec{x}_j)$.
  • Each $f_j$ is $\frac{G}{n}$-Lipschitz (i.e., $\|\vec{\nabla} f_j(\vec{\theta})\|_2 \leq \frac{G}{n}$ for all $\vec{\theta}$).
  • What does this imply about how Lipschitz $f$ is?
• Initialize with $\theta^{(1)}$ satisfying $\|\vec{\theta}^{(1)} - \vec{\theta}^*\|_2 \leq R$.

Stochastic Gradient Descent

• Pick some initial $\vec{\theta}^{(1)}$.
• Set step size $\eta = \frac{R}{G\sqrt{t}}$.
• For $i = 1, \ldots, t$
  • Pick random $j_i \in 1, \ldots, n$.
  • $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \cdot \vec{\nabla} f_{j_i}(\vec{\theta}^{(i)})$
• Return $\hat{\vec{\theta}} = \frac{1}{t} \sum_{i=1}^{t} \vec{\theta}^{(i)}$. 
\[ \tilde{\theta}^{(i+1)} = \tilde{\theta}^{(i)} - \eta \cdot \nabla f_j(\tilde{\theta}^{(i)}) \] vs. \[ \hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} - \eta \cdot \nabla f(\hat{\theta}^{(i)}) \]

**Note that:** \[ \mathbb{E}[\nabla f_j(\tilde{\theta}^{(i)})] = \frac{1}{n} \nabla f(\hat{\theta}^{(i)}). \]

Analysis extends to any algorithm that takes the gradient step in expectation (minibatch SGD, randomly quantized, measurement noise, differentially private, etc.)
Theorem – SGD on Convex Lipschitz Functions: SGD 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 $\theta^*$, outputs $\hat{\theta}$ satisfying: $\mathbb{E}[f(\hat{\theta})] \leq f(\theta^*) + \epsilon$. 
Theorem – SGD on Convex Lipschitz Functions: SGD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius $R$ of $\theta^*$, outputs $\hat{\theta}$ satisfying: $\mathbb{E}[f(\hat{\theta})] \leq f(\theta^*) + \epsilon$.

Step 1: $f(\hat{\theta}) - f(\theta^*) \leq \frac{1}{t} \sum_{i=1}^{t} [f(\theta^{(i)}) - f(\theta^*)]$
**Theorem – SGD on Convex Lipschitz Functions:** SGD run with \( t \geq \frac{R^2G^2}{\epsilon^2} \) iterations, \( \eta = \frac{R}{G\sqrt{t}} \), and starting point within radius \( R \) of \( \theta^* \), outputs \( \hat{\theta} \) satisfying: \( \mathbb{E}[f(\hat{\theta})] \leq f(\theta^*) + \epsilon \).

**Step 1:** \( f(\hat{\theta}) - f(\theta^*) \leq \frac{1}{t} \sum_{i=1}^{t} [f(\theta^{(i)})) - f(\theta^*)] \)

**Step 2:** \( \mathbb{E}[f(\hat{\theta}) - f(\theta^*)] \leq \frac{n}{t} \cdot \mathbb{E} \left[ \sum_{i=1}^{t} [f_{ji}(\theta^{(i)}) - f_{ji}(\theta^*)] \right] \).
Theorem – SGD on Convex Lipschitz Functions: SGD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius $R$ of $\theta^*$, outputs $\hat{\theta}$ satisfying: $\mathbb{E}[f(\hat{\theta})] \leq f(\theta^*) + \epsilon$.

**Step 1:** $f(\hat{\theta}) - f(\theta^*) \leq \frac{1}{t} \sum_{i=1}^{t} [f(\theta^{(i)}) - f(\theta^*)]$

**Step 2:** $\mathbb{E}[f(\hat{\theta}) - f(\theta^*)] \leq \frac{n}{t} \cdot \mathbb{E} \left[ \sum_{i=1}^{t} [f_j(\theta^{(i)}) - f_j(\theta^*)] \right]$.

**Step 3:** $\mathbb{E}[f(\hat{\theta}) - f(\theta^*)] \leq \frac{n}{t} \cdot \mathbb{E} \left[ \sum_{i=1}^{t} [f_j(\theta^{(i)}) - f_j(\theta^{\text{off}})] \right]$. 
**Theorem – SGD on Convex Lipschitz Functions:** SGD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius $R$ of $\theta^*$, outputs $\hat{\theta}$ satisfying: $\mathbb{E}[f(\hat{\theta})] \leq f(\theta^*) + \epsilon$.

**Step 1:** $f(\hat{\theta}) - f(\theta^*) \leq \frac{1}{t} \sum_{i=1}^{t}[f(\theta^{(i)}) - f(\theta^*)]$ 

**Step 2:** $\mathbb{E}[f(\hat{\theta}) - f(\theta^*)] \leq \frac{n}{t} \cdot \mathbb{E} \left[ \sum_{i=1}^{t}[f_{ji}(\theta^{(i)}) - f_{ji}(\theta^*)] \right]$.

**Step 3:** $\mathbb{E}[f(\hat{\theta}) - f(\theta^*)] \leq \frac{n}{t} \cdot \mathbb{E} \left[ \sum_{i=1}^{t}[f_{ji}(\theta^{(i)}) - f_{ji}(\theta^{\text{off}})] \right]$.

**Step 4:** $\mathbb{E}[f(\hat{\theta}) - f(\theta^*)] \leq \frac{n}{t} \cdot R \cdot \frac{G}{n} \cdot \sqrt{t} = \frac{RG}{\sqrt{t}}$. 

**OGD bound**
Stochastic gradient descent generally makes more iterations than gradient descent.

Each iteration is much cheaper (by a factor of $n$).

\[ \nabla \sum_{j=1}^{n} f_j(\theta) \text{ vs. } \nabla f_j(\theta) \]
When $f(\vec{\theta}) = \sum_{j=1}^{n} f_j(\vec{\theta})$ and $\|\vec{\nabla} f_j(\vec{\theta})\|_2 \leq \frac{G}{n}$:

**Theorem – SGD:** After $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations outputs $\hat{\theta}$ satisfying:

$$\mathbb{E}[f(\hat{\theta})] \leq f(\theta^*) + \epsilon.$$

When $\|\vec{\nabla} f(\vec{\theta})\|_2 \leq \bar{G}$:

**Theorem – GD:** After $t \geq \frac{R^2 \bar{G}^2}{\epsilon^2}$ iterations outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta^*) + \epsilon.$$
• Introduced the online optimization problem and the notion of regret to measure error in this setting.

• Introduced online gradient descent, which can solve online convex optimization with average regret approaching 0.

• Introduced stochastic gradient descent, an offline optimization method that can be analyzed as a special case of online gradient descent.