Proximal operator and methods

Master 2 Data Science, Univ. Paris Saclay

Robert M. Gower
A Datum Function

\[ f_i(w) := \ell (h_w(x^i), y^i) + \lambda R(w) \]

\[
\frac{1}{n} \sum_{i=1}^{n} \ell (h_w(x^i), y^i) + \lambda R(w) = \frac{1}{n} \sum_{i=1}^{n} (\ell (h_w(x^i), y^i) + \lambda R(w)) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)
\]

Finite Sum Training Problem

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)
\]
The Training Problem

Solving the training problem:

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)
\]

Reference method: Gradient descent

\[
\nabla \left( \frac{1}{n} \sum_{i=1}^{n} f_i(w) \right) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w)
\]

Gradient Descent Algorithm

Set \( w^1 = 0 \), choose \( \alpha > 0 \).

for \( t = 1, 2, 3, \ldots, T \)

\[
w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(w^t)
\]

Output \( w^{T+1} \)
Convergence GD I

**Theorem**

Let $f$ be convex and $L$-smooth.

$$f(w^T) - f(w^*) \leq \frac{2L||w^1 - w^*||^2}{T - 1} = O \left( \frac{1}{T} \right).$$

Where

$$w^{t+1} = w^t - \frac{1}{L} \nabla f(w^t)$$

Proof on board

$$\Rightarrow \text{ for } \frac{f(w^T) - f(w^*)}{||w^1 - w^*||^2} \leq \epsilon \text{ we need } T \geq \frac{2L}{\epsilon} = O \left( \frac{1}{\epsilon} \right)$$
Convergence GD I

**Theorem**

Let $f$ be convex and $L$-smooth.

$$f(w^T) - f(w^*) \leq \frac{2L||w^1 - w^*||_2^2}{T - 1} = O\left(\frac{1}{T}\right).$$

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

Proof on board

Is $f$ always differentiable?

$$\Rightarrow \quad \text{for} \quad \frac{f(w^T) - f(w^*)}{||w^1 - w^*||_2^2} \leq \epsilon \quad \text{we need} \quad T \geq \frac{2L}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$$
Convergence GD I

**Theorem**

Let $f$ be convex and $L$-smooth.

$$f(w^T) - f(w^*) \leq \frac{2L\|w^1 - w^*\|_2^2}{T - 1} = O\left(\frac{1}{T}\right).$$

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

Proof on board

Not true for many problems

Is $f$ always differentiable?

$$\Rightarrow \text{for } \frac{f(w^T) - f(w^*)}{\|w^1 - w^*\|_2^2} \leq \epsilon \text{ we need } T \geq \frac{2L}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$$
Change notation: Keep loss and regularizor separate

**Loss function**

\[ L(w) := \frac{1}{n} \sum_{i=1}^{n} \ell \left( h_w(x^i), y^i \right) \]

**The Training problem**

\[ \min_w L(w) + \lambda R(w) \]

- If \( L \) or \( R \) is not differentiable, \( L+R \) is not differentiable.
- If \( L \) or \( R \) is not smooth, \( L+R \) is not smooth.
Non-smooth Example

\[ L(w) + R(w) = \frac{1}{2} \|w\|_2^2 + \|w\|_1 \]
Non-smooth Example

\[ L(w) + R(w) = \frac{1}{2} \|w\|^2_2 + \|w\|_1 \]
Non-smooth Example

\[ L(w) + R(w) = \frac{1}{2} \|w\|_2^2 + \|w\|_1 \]

Does not fit.
Not smooth

Need more tools
Non-smooth Example

\[ L(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1 \]

Does not fit.
Not smooth

\[ f(w) + \langle g, y - w \rangle \]

Need more tools
Convexity: Subgradient

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex

$$\partial f(w) := \{g \in \mathbb{R}^n : f(y) \geq f(w) + \langle g, y - w \rangle, \forall y \in \text{dom}(f)\}$$
Convexity: Subgradient

Let $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ be convex

$$\partial f(w) := \{g \in \mathbb{R}^n : f(y) \geq f(w) + \langle g, y - w \rangle, \forall y \in \text{dom}(f)\}$$

$f(w) + \langle g, y - w \rangle$

$g = 0$

$w^* = \arg\min_w f(w) \iff 0 \in \partial f(w^*)$
Examples: L1 norm

\[
\partial |w| = \begin{cases} 
-1 & \text{if } w < 0 \\
[-1, 1] & \text{if } w = 0 \\
1 & \text{if } w > 0 
\end{cases}
\]

\[
\partial ||w||_1 = (\partial |w_1|, \ldots, \partial |w_d|)
\]

\[
|w| + \left< \frac{1}{2}, y - w \right>
\]
Examples

Lasso

$$\min_{w \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} (y^i - \langle w, a^i \rangle)^2 + \lambda \|w\|_1$$

Low Rank Matrix Recovery

$$\min_{W \in \mathbb{R}^{d \times d}} \frac{1}{n} \sum_{i=1}^{n} \|AW - Y\|_F^2 + \lambda \|W\|_*$$

SVM with soft margin

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y^i \langle w, a^i \rangle\} + \lambda \|w\|_2^2$$
Assumptions for this class

The Training problem

$$\min_w L(w) + \lambda R(w)$$

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

$R(w)$ is convex and “easy to optimize”

What does this mean?
Optimality conditions

The Training problem

\[ w^* = \arg \min_{w \in \mathbb{R}^d} L(w) + \lambda R(w) \]

\[ 0 \in \partial (L(w^*) + \lambda R(w^*)) = \nabla L(w^*) + \lambda \partial R(w^*) \]

\[-\nabla L(w^*) \in \lambda \partial R(w^*) \]
Working example: Lasso

Lasso

\[
\min_{w \in \mathbb{R}^d} \frac{1}{2n} \| Aw - y \|_2^2 + \lambda \| w \|_1
\]

\[
A = [a^1, \ldots, a^n]^\top \Rightarrow \sum_{i=1}^{n} (y^i - \langle w, a^i \rangle)^2 = \| Aw - y \|_2^2
\]

\[-\nabla L(w^*) \in \partial R(w^*) \quad \text{arrow} \quad -A^\top (Aw^* - y) \in \partial \| w^* \|_1\]

Difficult inclusion, do iteratively.
Using $\mathcal{L}$-smoothness of $L$:

$$L(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The $w$ that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{\mathcal{L}} \nabla L(y)$$

$$\lambda R(w)$$
Proximal method I

Using $\mathcal{L}$-smoothness of $L$:

$$L(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} \|w - y\|^2, \quad \forall w, y \in \mathbb{R}^d$$

The $w$ that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{\mathcal{L}} \nabla L(y)$$

But what about $R(w)$? Adding on $+ \lambda R(w)$ to upper bound:
Proximal method I

Using $\mathcal{L}$–smoothness of $L$:

$$L(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The $w$ that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{\mathcal{L}} \nabla L(y)$$

But what about $R(w)$? Adding on $+\lambda R(w)$ to upper bound:

$$L(w) + \lambda R(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2 + \lambda R(w)$$
Proximal method I

Using $\mathcal{L}$-smoothness of $L$:

$$L(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} \| w - y \|^2, \quad \forall w, y \in \mathbb{R}^d$$

The $w$ that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{\mathcal{L}} \nabla L(y)$$

But what about $R(w)$? Adding on $+ \lambda R(w)$ to upper bound:

$$L(w) + \lambda R(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} \| w - y \|^2 + \lambda R(w)$$

Can we minimize the right-hand side?
Proximal method II

Minimizing the right-hand side of

\[ L(w) + \lambda R(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} \|w - y\|^2 + \lambda R(w) \]

\[ \arg \min_w L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} \|w - y\|^2 + \lambda R(w) \]

\[ = \arg \min_w \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} \|w - y\|^2 + \lambda R(w) \]

\[ = \arg \min_w \frac{1}{2} \|w - (y - \frac{1}{\mathcal{L}} \nabla L(y))\|^2 + \frac{\lambda}{\mathcal{L}} R(w) \]

\[ \text{prox}_{\lambda \frac{\mathcal{L}}{R} R}(v) := \arg \min_w \frac{1}{2} \|w - v\|^2 + \frac{\lambda}{\mathcal{L}} R(w) \]

What is this prox operator?
Proximal Operator I

Let $f(w)$ be convex. We define the proximal operator as

$$\text{prox}_f(v) := \arg \min_w \frac{1}{2}||w - v||^2_2 + f(w)$$

Let $w_v = \text{prox}_f(v)$. Using optimality conditions

$$0 \in \partial \left( \frac{1}{2}||w_v - v||^2_2 + f(w) \right) = w_v - v + \partial f(w_v)$$

Rearranging

$$\text{prox}_f(v) = w_v \in v - \partial f(w_v)$$
Proximal Operator II

\[ \text{prox}_f(v) := \arg \min_w \frac{1}{2} \| w - v \|_2^2 + f(w) \]

**Exe:**

1) If \( f(w) = \sum_{i=1}^{d} f_i(w_i) \) then

\[ \text{prox}_f(v) = (\text{prox}_{f_1}(v_1), \ldots, \text{prox}_{f_d}(v_d)) \]

2) If \( f(w) = I_C(w) := \begin{cases} 
0 & \text{if } w \in C \\
\infty & \text{if } w \notin C
\end{cases} \) where \( C \) is closed and convex then \( \text{prox}_f(v) = \text{proj}_C(v) \)

3) If \( f(w) = \langle b, w \rangle + c \) then \( \text{prox}_f(v) = v - b \)

4) If \( f(w) = \frac{\lambda}{2} w^\top A w + \langle b, w \rangle \) where \( A > 0, A = A^\top \) then

\[ \text{prox}_{\lambda f}(v) = (I + \lambda A)^{-1}(v - b) \]
Proximal Operator III: Soft thresholding

\[
\text{prox}_{\lambda \Vert w \Vert_1}(v) := \arg \min_w \frac{1}{2} \Vert w - v \Vert_2^2 + \lambda \Vert w \Vert_1
\]

**Exe:**

1) Let \( \alpha \in \mathbb{R} \). If \( \alpha^* = \arg \min_{\alpha} \frac{1}{2} (\alpha - v)^2 + \lambda |\alpha| \) then

\[
\alpha^* \in v - \lambda \partial |\alpha^*| \quad (I)
\]

2) If \( \lambda < v \) show \((I)\) gives \( \alpha^* = v - \lambda \)

3) If \( v < -\lambda \) show \((I)\) gives \( \alpha^* = v + \lambda \)

4) Show that

\[
\text{prox}_{\lambda \Vert \alpha \Vert_1}(v) = \begin{cases} 
v - \lambda & \text{if } \lambda < v \\
0 & \text{if } -\lambda \leq v \leq \lambda \\
v + \lambda & \text{if } v < -\lambda.
\end{cases}
\]
Proximal Operator IV: Singular value thresholding

\[ S_\lambda(v) := \arg \min_w \frac{1}{2} \|w - v\|_2^2 + \lambda \|w\|_1 \]

Similarly, the prox of the nuclear norm for matrices:

\[ U \text{diag}(S_\lambda(\text{diag}(\sigma(A))))V^\top := \arg \min_{W \in \mathbb{R}^{d \times d}} \frac{1}{2} \|W - A\|_F^2 + \lambda \|W\|_* \]

where \( A = U \text{diag}(\sigma(A))V^\top \) is a SVD decomposition.
Proximal method V

Minimizing the right-hand side of

\[
L(w) + \lambda R(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} \|w - y\|^2 + \lambda R(w)
\]

\[
\arg \min_w L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} \|w - y\|^2 + \lambda R(w)
\]

\[
= \arg \min_w \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} \|w - y\|^2 + \lambda R(w)
\]

\[
= \arg \min_w \frac{1}{2} \|w - (y - \frac{1}{\mathcal{L}} \nabla L(y))\|^2 + \frac{\lambda}{\mathcal{L}} R(w)
\]

\[
= \text{prox}_{\frac{\lambda}{\mathcal{L}} R} \left( y - \frac{1}{\mathcal{L}} \nabla L(w) \right)
\]

Make iterative method based on this upper bound minimization
The Proximal Gradient Method

Solving the training problem:
\[ \min_w L(w) + \lambda R(w) \]

- \( L(w) \) is differentiable, \( \mathcal{L} \)-smooth and convex
- \( R(w) \) is convex and prox friendly

**Proximal Gradient Descent**

Set \( w^1 = 0 \).

for \( t = 1, 2, 3, \ldots, T \)
\[
    w^{t+1} = \text{prox}_{\lambda R / \mathcal{L}} \left( w^t - \frac{1}{\mathcal{L}} \nabla L(w^t) \right)
\]

Output \( w^{T+1} \)
The Training problem

$$\min_w L(w) + \lambda R(w)$$

$$-\nabla L(w^*) \in \lambda \partial R(w^*)$$

$$w^* + \gamma \nabla L(w^*) \in w^* - (\lambda \gamma) \partial R(w^*)$$

$$w^* \in (w^* - \gamma \nabla L(w^*)) - (\lambda \gamma) \partial R(w^*)$$

$$\text{prox}_f(v) = w_v \in v - \partial f(w_v)$$

$$w^* = \text{prox}_{\lambda \gamma R}(w^* - \gamma \nabla L(w^*))$$

Optimal is a fixed point.
Working example: Lasso

Lasso

$$\min_{w \in \mathbb{R}^d} \frac{1}{2n} \| Aw - y \|_2^2 + \lambda \| w \|_1$$

$$A = [a^1, \ldots, a^n]^\top \Rightarrow \sum_{i=1}^{n} (y^i - \langle w, a^i \rangle)^2 = \| Aw - y \|_2^2$$

$$w^{t+1} = \operatorname{prox}_{\lambda \| w \|_1 / \mathcal{L}} \left( w^t - \frac{1}{2n\mathcal{L}} A^\top (Aw^t - y) \right)$$

$$= S_{\lambda / \mathcal{L}} \left( w^t - \frac{1}{2\sigma_{\text{max}}(A)^2} A^\top (Aw^t - y) \right)$$

$$\mathcal{L} = \frac{\sigma_{\text{max}}(A)^2}{n}$$

Working example: Lasso

\[
\begin{align*}
\text{Lasso} \\
\min_{w \in \mathbb{R}^d} \frac{1}{2n} \| A w - y \|^2_2 + \lambda \| w \|_1
\end{align*}
\]

\[A = [a^1, \ldots, a^n]^\top \Rightarrow \sum_{i=1}^n (y^i - \langle w, a^i \rangle)^2 = \| A w - y \|^2_2\]

\[
\begin{align*}
\omega^{t+1} &= \text{prox}_{\frac{\lambda \| w \|_1}{\mathcal{L}}} \left( w^t - \frac{1}{2n \mathcal{L}} A^\top (A w^t - y) \right) \\
&= S_{\frac{\lambda}{\mathcal{L}}} \left( w^t - \frac{1}{2\sigma_{\max}(A)^2} A^\top (A w^t - y) \right)
\end{align*}
\]

\[\mathcal{L} = \frac{\sigma_{\max}(A)^2}{n}\]
Convergence of Prox-GD

**Theorem (Beck Teboulle 2009)**

Let \( f(w) = L(w) + \lambda R(w) \) where

\[ L(w) \] is differentiable, \( \mathcal{L} \)-smooth and convex

\[ R(w) \] is convex and prox friendly

Then

\[
f(w^T) - f(w^*) \leq \frac{L\|w^1 - w^*\|_2^2}{2T} = O\left(\frac{1}{T}\right).
\]

where

\[
w^{t+1} = w^{t+1} = \text{prox}_{\lambda R/\mathcal{L}} \left( w^t - \frac{1}{\mathcal{L}} \nabla L(w^t) \right)
\]
Convergence of Prox-GD

Theorem (Beck Teboulle 2009)

Let \( f(w) = L(w) + \lambda R(w) \) where

- \( L(w) \) is differentiable, \( \mathcal{L} \)-smooth and convex
- \( R(w) \) is convex and prox friendly

Then

\[
f(w^T) - f(w^*) \leq \frac{L\|w^1 - w^*\|^2_2}{2T} = O\left(\frac{1}{T}\right).
\]

where

\[
w^{t+1} = \text{prox}_{\lambda R/\mathcal{L}} \left( w^t - \frac{1}{\mathcal{L}} \nabla L(w^t) \right)
\]
The FISTA Method

Solving the *training problem*: \[
\min_{w} L(w) + \lambda R(w)
\]

The FISTA Algorithm

Set \( w^1 = 0 = z^1, \beta^1 = 1 \)
for \( t = 1, 2, 3, \ldots, T \)

\[
w^{t+1} = \text{prox}_{\frac{\lambda R}{\mathcal{L}}} \left( z^t - \frac{1}{\mathcal{L}} \nabla L(z^t) \right)
\]

\[
\beta^{t+1} = \frac{1 + \sqrt{1 + 4(\beta^t)^2}}{2 \beta^t - 1}
\]

\[
z^{t+1} = w^{t+1} + \frac{\beta^t - 1}{\beta^{t+1}} (w^{t+1} - w^t)
\]

Output \( w^{T+1} \)
The FISTA Method

Solving the training problem:

\[
\min_w L(w) + \lambda R(w)
\]

---

**The FISTA Algorithm**

Set \( w^1 = 0 = z^1, \beta^1 = 1 \)

for \( t = 1, 2, 3, \ldots, T \)

\[
\begin{align*}
w^{t+1} &= \text{prox}_{\lambda R/\mathcal{L}} \left( z^t - \frac{1}{\mathcal{L}} \nabla L(z^t) \right) \\
\beta^{t+1} &= \frac{1 + \sqrt{1 + 4(\beta^t)^2}}{2} \\
z^{t+1} &= w^{t+1} + \frac{\beta^t - 1}{\beta^{t+1}} (w^{t+1} - w^t)
\end{align*}
\]

Output \( w^{T+1} \)

---

Weird, but it works
Convergence of FISTA

Theorem (Beck Teboulle 2009)

Let \( f(w) = L(w) + \lambda R(w) \) where

- \( L(w) \) is differentiable, \( \mathcal{L} \)-smooth and convex
- \( R(w) \) is convex and prox friendly

Then

\[
 f(w^T) - f(w^*) \leq \frac{2L\|w^1 - w^*\|_2^2}{(T + 1)^2} = O \left( \frac{1}{T^2} \right).
\]

Where \( w^t \) are given by the FISTA algorithm
Convergence of FISTA

Theorem (Beck Teboulle 2009)

Let $f(w) = L(w) + \lambda R(w)$ where

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

$R(w)$ is convex and prox friendly

Then

$$f(w^T) - f(w^*) \leq \frac{2L||w^1 - w^*||_2^2}{(T+1)^2} = O \left( \frac{1}{T^2} \right).$$

Where $w^t$ are given by the FISTA algorithm

Convergence lower bounds

Theorem (Nesterov)

There exists a function $f(w)$ that is $L$–smooth and convex such that for any optimization algorithm where

$$w^{t+1} \in w^t + \text{span} \left( \nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t) \right)$$

Then

$$\min_{i=1,\ldots,T} f(w^i) - f(w^*) \geq \frac{3L\|w^1 - w^*\|_2^2}{32(T + 1)^2} = O \left( \frac{1}{T^2} \right).$$

Where $w^t$ are given by the FISTA algorithm and $T \leq \frac{d - 1}{2}$. 

Convergence lower bounds

**Theorem (Nesterov)**

There exists a function $f(w)$ that is $L$–smooth and convex such that for any optimization algorithm where

$$w^{t+1} \in w^t + \text{span} \left( \nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t) \right)$$

Then

$$\min_{i=1, \ldots, T} f(w^i) - f(w^*) \geq \frac{3L\|w^1 - w^*\|_2^2}{32(T+1)^2} = O \left( \frac{1}{T^2} \right).$$

Where $w^t$ are given by the FISTA algorithm and $T \leq \frac{d - 1}{2}$. 

---

Lab Session 02/10

Bring your laptop!
Lab Session 02/10

Bring your laptop!
Introduction to Stochastic Gradient Descent
Optimization Sum of Terms

A Datum Function
\[ f_i(w) := \ell \left( h_w(x^i), y^i \right) + \lambda R(w) \]

\[
\frac{1}{n} \sum_{i=1}^{n} \ell \left( h_w(x^i), y^i \right) + \lambda R(w) = \frac{1}{n} \sum_{i=1}^{n} \left( \ell \left( h_w(x^i), y^i \right) + \lambda R(w) \right)
\]
\[ = \frac{1}{n} \sum_{i=1}^{n} f_i(w) \]

Finite Sum Training Problem
\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) =: f(w)
\]
The Training Problem

Solving the training problem:

\[ \min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \]

Reference method: Gradient descent

\[ \nabla \left( \frac{1}{n} \sum_{i=1}^{n} f_i(w) \right) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w) \]

Gradient Descent Algorithm

Set \( w^0 = 0 \), choose \( \alpha > 0 \).

for \( t = 1, 2, 3, \ldots, T \)

\[ w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(w^t) \]

Output \( w^{T+1} \)
The Training Problem

Solving the training problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Problem with Gradient Descent:
Each iteration requires computing a gradient $\nabla f_i(w)$ for each data point. One gradient for each cat on the internet!

Gradient Descent Algorithm

Set $w^0 = 0$, choose $\alpha > 0$.
for $t = 1, 2, 3, \ldots, T$

$$w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(w^t)$$

Output $w^{T+1}$
Is it possible to design a method that uses only the gradient of a single data function $f_i(w)$ at each iteration?
Stochastic Gradient Descent

Is it possible to design a method that uses only the gradient of a single data function $f_i(w)$ at each iteration?

**Unbiased Estimate**
Let $j$ be a random index sampled from $\{1, \ldots, n\}$ selected uniformly at random. Then

$$\mathbb{E}_j \left[ \nabla f_j(w) \right] = \frac{1}{n} \sum \nabla f_i(w) = \nabla f(w)$$
Is it possible to design a method that uses only the gradient of a **single** data function $f_i(w)$ at each iteration?

**Unbiased Estimate**

Let $j$ be a random index sampled from $\{1, \ldots, n\}$ selected uniformly at random. Then

$$
\mathbb{E}_j [\nabla f_j(w)] = \frac{1}{n} \sum \nabla f_i(w) = \nabla f(w)
$$

Use $\nabla f_j(w) \approx \nabla f(w)$
Stochastic Gradient Descent

**Stochastic Gradient Descent Algorithm**

Set $w^0 = 0$, choose $\alpha > 0$.

for $t = 1, 2, 3, \ldots, T$

Sample $j \in \{1, \ldots, n\}$

$w^{t+1} = w^t - \alpha \nabla f_j(w^t)$

Output $w^{T+1}$