COS324: Introduction to Machine Learning

Lecture 10: Gradient Methods in Machine Learning

Prof. Elad Hazan & Prof. Yoram Singer

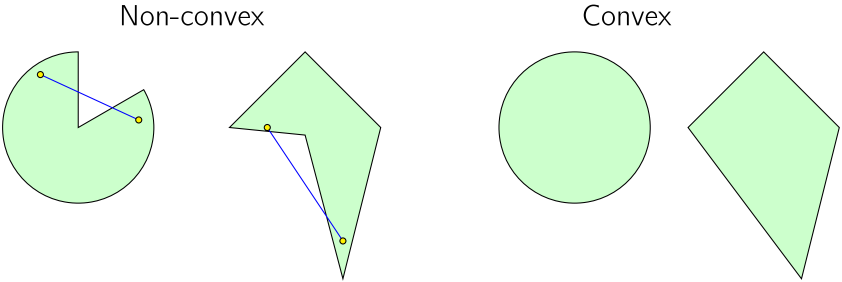
Recap & Today

- Reminder of convexity, GD, and SGD
- Linear regression
  1. Problem definition
  2. Direct solution
  3. SGD for linear regression
- Binary classification
  1. Surrogate losses
  2. Sub-gradients
  3. Perceptron revisited
  4. SGD for binary classification
- Beyond binary learning problems

Convex Sets

Ω is convex set: $\forall\, \mathbf{u}, \mathbf{v} \in \Omega$, line segment between $\mathbf{u}$ and $\mathbf{v}$ is in Ω

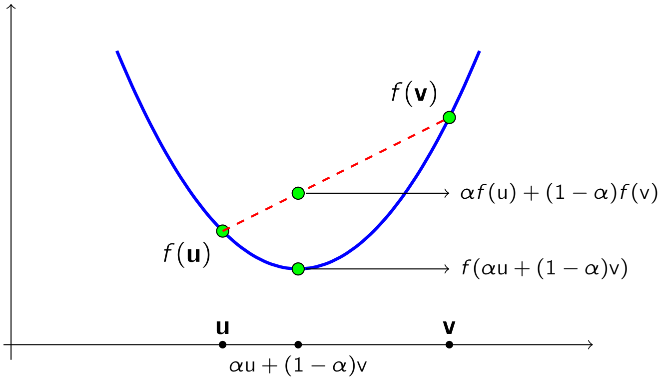
$$\forall\, \alpha \in [0, 1] \;\; \alpha\mathbf{u} + (1-\alpha)\mathbf{v} \in \Omega$$



Convex Functions

Function $f : \Omega \to \mathbb{R}$ is convex if $\forall \mathbf{u}, \mathbf{v} \in C$ and $\alpha \in [0, 1]$,

$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \;\leq\; \alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v})$$

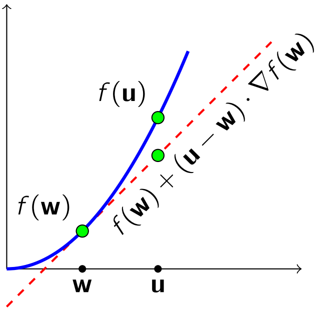


Tangents Lie Below f

Gradient of f at $\mathbf{w}$: $\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathrm{w})}{\partial w_1}, \ldots, \frac{\partial f(\mathrm{w})}{\partial w_d}\right)$

If f is convex and differentiable, then

$$\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \nabla f(\mathbf{w}) \cdot (\mathbf{u} - \mathbf{w})$$



Convex Optimization & Learning

Convex optimization,

$$\min_{w \in \Omega} f(\mathbf{w})$$

where f is a convex function and Ω is a convex set

C.O. for Machine learning,

$$f(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \ell(\mathbf{w}, (\mathbf{x}_i, y_i))$$

where $\ell()$ is a convex loss function in $\mathbf{w}$ and assume $\Omega = \mathbb{R}^d$

Often abbreviate $f_i(\mathbf{w}) \stackrel{\text{def}}{=} \ell(\mathbf{w}, (\mathbf{x}_i, y_i))$ or $\ell_i(\mathbf{w}) \stackrel{\text{def}}{=} \ell(\mathbf{w}, (\mathbf{x}_i, y_i))$

Gradient Descent

- Initialize $\mathbf{w}^1$ (typically $\mathbf{w}^1 = \mathbf{0}$)
- For $t = 1, \ldots, T$:
  - Set learning-rate η^t (often fixed)
  - Perform gradient descent step:

$$\begin{aligned} \mathbf{w}^{t+1} &= \mathbf{w}^t - \eta^t \nabla f(\mathbf{w}^t) \\ &= \mathbf{w}^t - \eta^t \frac{1}{|S|} \sum_{i \in S} \nabla f_i(\mathbf{w}^t) \end{aligned}$$

- Output $\bar{\mathbf{w}}^T = \frac{1}{T} \sum_{t=1}^{t} \mathbf{w}^t$

Gradient Descent - Properties

- Assume or constrain $\|\mathbf{w}\| \leq D/2$ therefore

$$\Rightarrow \; \|\mathbf{w}^t - \mathbf{w}^\star\| \leq \|\mathbf{w}^t\| + \|\mathbf{w}^\star\| \leq D$$

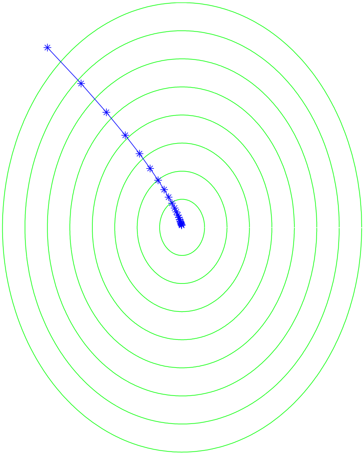
- Assume $\|\nabla f(\mathbf{w}^t)\| \leq G$

- Convergence rate of GD:

$$f(\bar{\mathbf{w}}^T) - f(\mathbf{w}^\star) \leq \frac{DG}{\sqrt{T}}$$

- However, each iteration requires $O(dm)$ operations
  [d – dimension, m – number of examples]

Iterates of Gradient Descent



Stochastic Gradient Descent

- Initialize $\mathbf{w}^1$ (typically $\mathbf{w}^1 = \mathbf{0}$)
- For $t = 1, \ldots, T$:
  - Set learning-rate η^t (typically decreasing)
  - Perform stochastic gradient descent step:
    - Choose $S' \subset S$ at random
    - Update

$$\begin{aligned} \mathbf{w}^{t+1} &= \mathbf{w}^t - \eta^t \nabla \hat{f}(\mathbf{w}^t) \\ &= \mathbf{w}^t - \eta^t \frac{1}{|S'|} \sum_{i \in S'} \nabla f_i(\mathbf{w}^t) \end{aligned}$$

- Output $\bar{\mathbf{w}}^T = \frac{1}{T} \sum_{t=1}^{t} \mathbf{w}^t$

Stochastic Gradient Descent - Properties

- Assume that

$$\forall i : \ \|\nabla f_i(\mathbf{w}^t)\| \le G$$

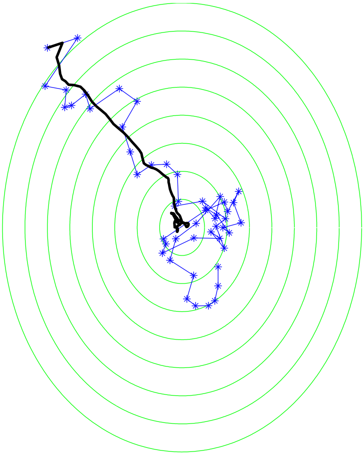
in contrast to GD, $\|\nabla f(\mathbf{w}^t)\| \le G$

- Convergence rate of GD:

$$\mathbb{E}\big[f(\bar{\mathbf{w}}^T)\big] - f(\mathbf{w}^\star) \le \frac{DG}{\sqrt{T}}$$

- Each iteration requires $O(dc)$ operations, c is sub-sample size

Iterates of SGD



- $*$ $f(w^t)$
- $\bullet$ $f\left(\frac{1}{t}\sum_{s\leq t} w^s\right)$

Regression Problems

- Automatic Kelly Blue Book: value assessment of used cars
- Collect sale information of cars: sold for \$\$
- For each car gather model year, # accidents, make, mileage # of previous owners, last sold for \$\$\$, ...

Year	Acci	Make	Mile	Ownr	Las\$	Cur\$
97	5	To	120	3	2.5	0.5
16	1	Te	17	0	80	60
12	0	Su	43	1	29	22
...	...	...	...	...	...	...
x_1	x_2	x_3	x_4	x_5	x_6	y

- How to represent symbolic features (Toyota, Tesla, Subaru) ?
- How to represent ordered sets (#accidents: $0 < 1 < 2 < \dots$) ?
- How to represent numeric features (v\$, $\log(v\$)$, $\log(v\$) - b$) ?

Linear Regression

- Each row is an example $\mathbf{x}_i \in \mathbb{R}^d$
- Last column is a target $y_i \in \mathbb{R}$
- Create $m \times d$ matrix s.t. $X_{i,j}$ is j'th entry of $\mathbf{x}_i$
- Create column vector $\mathbf{y}$ from $y_1, \ldots, y_n$
- Find a solution for the linear set of equations $X\mathbf{w} = \mathbf{y}$
  - Solution may not exist
  - Multiple solutions may exist
  - Complexity $O(md + d^3)$
- Approximately solve, $X\mathbf{w} \approx \mathbf{y}$ namely $\forall i: \ \mathbf{w} \cdot \mathbf{x}_i \approx y_i$
- Notion of $\approx$?

Regression Losses

- Convex loss $\ell : \mathbb{R} \to \mathbb{R}_+ \quad ; \quad \ell(z) = \ell(\mathbf{w} \cdot \mathbf{x} - y)$

- Example i induces convex loss

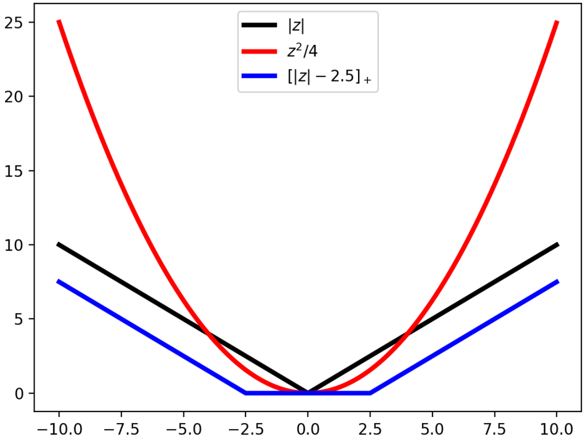
$$\ell_i(\mathbf{w}) = \ell(\mathbf{w} \cdot \mathbf{x}_i - y_i)$$

- Total loss:

$$f(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \ell\left(\mathbf{w} \cdot \mathbf{x}_i - y_i\right)$$

- Concrete losses $\ell(z) = \ldots$

$$z^2 \quad |z| \quad z^4 \quad \ldots \quad \min\{|z| - \gamma, 0\} \quad \exp(z) + \exp(-z)$$



|z|
z²/4
[|z| − 2.5]₊
25
20
15
10
5
0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Least Squares Regression $\ell(z) = \frac{1}{2}z^2$

- Parameters: radius D, learning rate η, number of iterations T
- Initialize: $\mathbf{w}^1 = \mathbf{0}$
- For $t = 1, \ldots, T$:
  - Choose $S' \subset S$ and calculate stochastic gradient

$$\nabla \hat{f}(\mathbf{w}^t) = \frac{1}{S'} \sum_{i \in S'} \underbrace{(\mathbf{w}^t \cdot \mathbf{x}_i - y_i)}_{\stackrel{\text{def}}{=} \Delta_i} \mathbf{x}_i$$

  - Update

$$\mathbf{w}^{t+\frac{1}{2}} = \mathbf{w}^t - \eta^t \nabla \hat{f}(\mathbf{w}^t)$$

$$\mathbf{w}^{t+1} = \min\left\{1, \frac{D}{\|\mathbf{w}^{t+\frac{1}{2}}\|}\right\} \mathbf{w}^{t+\frac{1}{2}}$$

- Output $\bar{\mathbf{w}}^T = \frac{1}{T} \sum_{t=1}^{t} \mathbf{w}^t$

Pesky Learning Rate

- Recall that $\eta = \frac{D}{G\sqrt{T}}$ where

$$\|\nabla f(\mathbf{w}^t)\| \leq G \qquad \|\mathbf{w}^t - \mathbf{w}^\star\| \leq D$$

- Assume or normalize such that $\forall i: \ \ \|\mathbf{x}_i\| \leq b \ \ |y_i| \leq c$

- Constrain $\forall t: \ \|\mathbf{w}^t\| \leq D/2$

- We thus get: $\quad \|\mathbf{w}^t - \mathbf{w}^\star\| \leq \|\mathbf{w}^t\| + \|\mathbf{w}^\star\| \leq D$

- In addition, we get a bound on gradients,

$$\begin{aligned}
\|(\mathbf{w}\cdot\mathbf{x}_i - y_i)\mathbf{x}_i\| &\leq |\mathbf{w}\cdot\mathbf{x}_i - y_i|\|\mathbf{x}_i\| \\
&\leq |\mathbf{w}\cdot\mathbf{x}_i - y_i| b \\
&\leq (|\mathbf{w}\cdot\mathbf{x}_i| + |y_i|) b \\
&\leq (Db + c) b \qquad\qquad [Cauchy-Schwarz]
\end{aligned}$$

- And we can set $\eta = \frac{D}{(Db+c)b\sqrt{T}}$ $\quad \ldots \quad$ but in practice ...

Binary Classification

- Examples $\mathbf{x}_i \in \mathbb{R}^d$
- Labels $y_i \in \{-1, +1\}$
- Predictor / classifier: $h_{\mathrm{w}}(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} - b)$
- b is called a bias term (assume it is zero for time being)
- Goal,

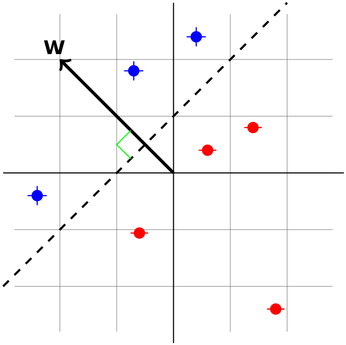
$$\min_{\mathrm{w}} \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}_i) \neq y_i]$$

- First attempt: define $z = y(\mathbf{w} \cdot \mathbf{x})$ and $\ell^{0-1}(z) = \mathbb{1}[z \leq 0]$
- Can we use (stochastic) gradient descent ?

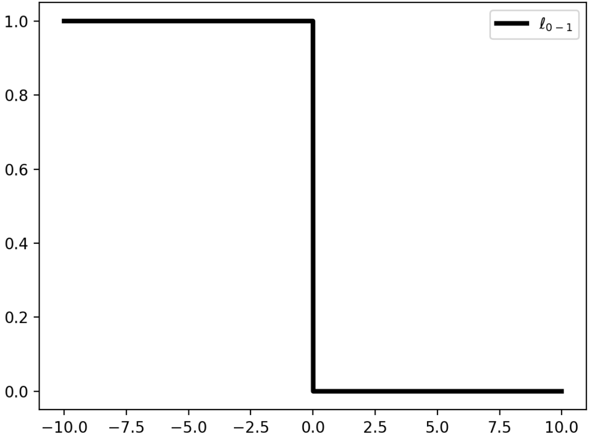
Linear Classifiers

- Domain: Euclidean space $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$
- Hypothesis class: thresholding linear predictors

$$h_{\mathrm{w}}(x) = \operatorname{sign}\left(\mathbf{w} \cdot \mathbf{x} - b\right)$$

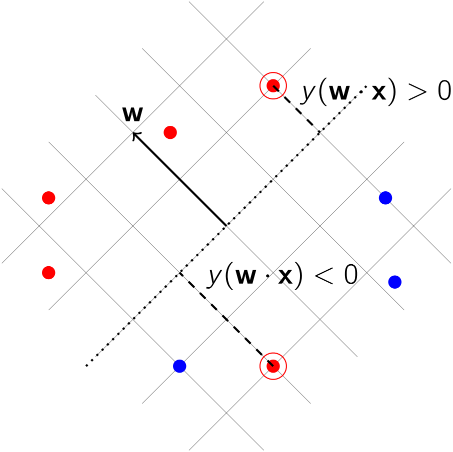


0-1 Loss



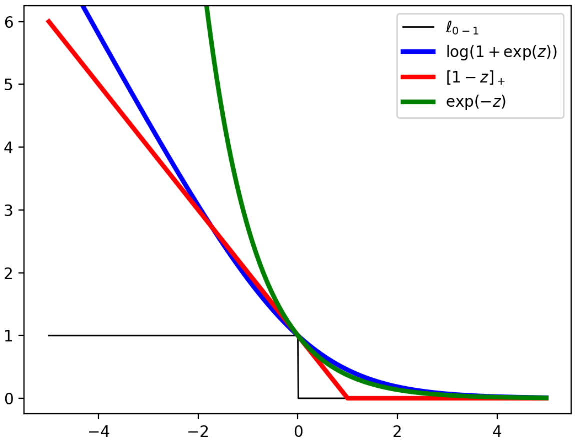
"Utopia": combinatorial problem which is NP-Hard

Classification Margin

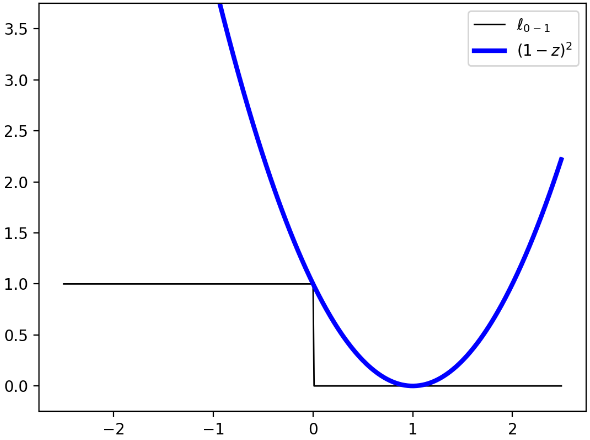


Surrogate Losses for Classification

- Convex losses w.r.t $z = y(\mathbf{w} \cdot \mathbf{x})$ which satisfy
$$\ell(z) \geq \ell^{0-1}(z)$$
- Exp-loss,
$$\exp(-z)$$
- Log-loss,
$$\log(1 + \exp(-z))$$
- Hinge-loss,
$$\max\{0, 1 - z\} = [1 - z]_+$$
- Squared-error with $\Delta = \mathbf{w} \cdot \mathbf{x} - y$,
$$\begin{aligned} \ell(\Delta) &= \Delta^2 = (\mathbf{w} \cdot \mathbf{x} - y)^2 \\ &= y^2(\mathbf{w} \cdot \mathbf{x} - y)^2 \\ &= (y(\mathbf{w} \cdot \mathbf{x}) - 1)^2 \quad \Rightarrow \quad \ell(z) = (1 - z)^2 \end{aligned}$$



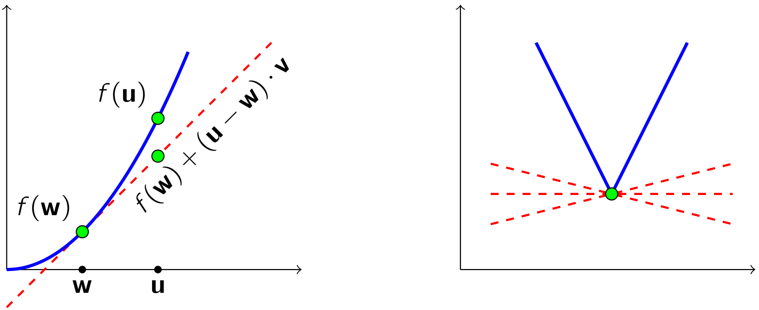
ℓ_{0-1}
$\log(1+\exp(z))$
$[1-z]_+$
$\exp(-z)$
6
5
4
3
2
1
0
−4
−2
0
2
4



ℓ_{0-1}
$(1-z)^2$
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−2
−1
0
1
2

Sub-gradients

- $\mathbf{v}$ is sub-gradient of f at $\mathbf{w}$ if $\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \mathbf{v} \cdot (\mathbf{u} - \mathbf{w})$
- The differential set, $\partial f(\mathbf{w})$, is the set of sub-gradients of f at $\mathbf{w}$
- Lemma: f is convex iff for every $\mathbf{w}$, $\partial f(\mathbf{w}) \neq \emptyset$

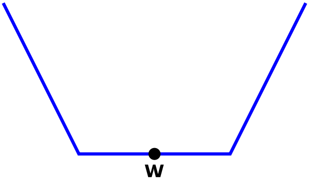


Optimality Property

f is "locally flat" around $\mathbf{w}$, i.e. $\mathbf{0}$ is a sub-gradient,

iff

$\mathbf{w}$ is a (not "the") global minimizer



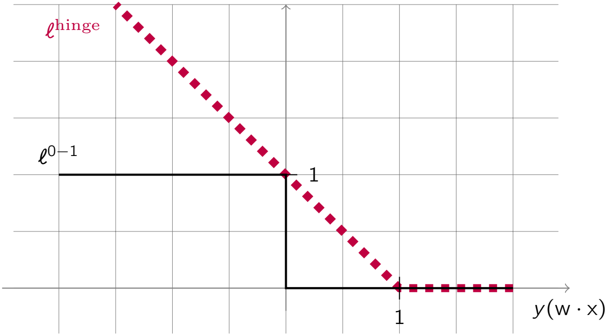
We can replace gradients with sub-gradients:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \mathbf{g}^t \quad \text{where} \quad \mathbf{g}^t \in \partial \hat{f}(\mathbf{w}^t)$$

Hinge Loss

$$\ell(z) = \max\{0, 1 - z\} = [1 - z]_+$$

$$\ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y)) \stackrel{\text{def}}{=} \max\{0, 1 - y(\mathbf{w} \cdot \mathbf{x})\}$$



Non-differentiable at $z = 1$

Can we use SGD ?

SGD for Hinge-Loss

- Fully stochastic case – single example
- Subgardient of $[1 - z]_+$,

$$\partial\ell(z) = \begin{cases} 0 & z > 1 \\ -1 & z < 1 \\ (-1, 0) & z = 1 \end{cases}$$

$$\partial\ell(\mathbf{w}, (\mathbf{x}, y)) = y\mathbf{x}\partial\ell(z) \;\; \text{where} \;\; z = y(\mathbf{w} \cdot \mathbf{x})$$

- SGD update on iteration t:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta\mathbf{g}^t \;\; \text{where} \;\; \mathbf{g}^t \in \partial\ell_t(\mathbf{w}^t)$$

$$\mathbf{w}^{t+1} = \begin{cases} \mathbf{w}^t + \eta y^t\mathbf{x}^t & y^t(\mathbf{w}^t \cdot \mathbf{x}^t) \leq 1 \\ \mathbf{w}^t & \text{otherwise} \end{cases}$$

SGD vs. Perceptron

- SGD

$$\mathbf{w}^{t+1} = \begin{cases} \mathbf{w}^t + \eta y^t \mathbf{x}^t & y^t(\mathbf{w}^t \cdot \mathbf{x}^t) \leq \mathbf{1} \\ \mathbf{w}^t & \text{otherwise} \end{cases}$$

- Perceptron

$$\mathbf{w}^{t+1} = \begin{cases} \mathbf{w}^t + \eta y^t \mathbf{x}^t & y^t(\mathbf{w}^t \cdot \mathbf{x}^t) \leq \mathbf{0} \\ \mathbf{w}^t & \text{otherwise} \end{cases}$$

SGD $\approx$ Perceptron

- Analysis of SGD assumes,

$$\|\nabla \ell_t(\mathbf{w}^t)\| \leq G \qquad \|\mathbf{w}^t - \mathbf{w}^\star\| \leq D$$

- Analysis of GD & SGD's implies,

$$\sum_{t=1}^{T} [1 - y_t(\mathbf{w}^t \cdot \mathbf{x}_t)]_+ \ \leq \ \sum_{t=1}^{T} [1 - y_t(\mathbf{w}^\star \cdot \mathbf{x}_t)]_+ \ + \ \sqrt{T} G D$$

- Analysis of Perecptron assumes,

$$\forall i : \|\mathbf{x}_i\| \leq 1 \ \ \exists \mathbf{w}^\star : \|\mathbf{w}^\star\| = 1 \ \wedge \ y_i(\mathbf{w}^\star \cdot \mathbf{x}_i) \geq \gamma$$

- Perceptron's mistake bound is,

$$\frac{1}{\gamma^2} \ \ \Rightarrow \ \ \sum_{t=1}^{T} \mathbb{1}\left[y_t(\mathbf{w}^t \cdot \mathbf{x}_t) \leq 0\right] \ \leq \ \frac{1}{\gamma^2}$$

SGD $\Rightarrow$ Perceptron

- Need to accommodate Perceptron's assumptions,

$$\forall i : \|\mathbf{x}_i\| \leq 1 \;\; \exists \mathbf{w}^\star : \|\mathbf{w}^\star\| = 1 \;\wedge\; y_i(\mathbf{w}^\star \cdot \mathbf{x}_i) \geq \gamma$$

- Constraining (by projecting) $\|\mathbf{w}^t\| \leq 1$ imply

$$\mathbf{w}^t \cdot \mathbf{x}_i \leq \|\mathbf{w}^t\| \, \|\mathbf{x}_i\| \leq 1$$

- Modify loss to be $[\gamma - y(\mathbf{w} \cdot \mathbf{x})]_+$

- We start at $\mathbf{w}^1 = \mathbf{0}$ & progress toward $\mathbf{w}^\star$ thus

$$\|\mathbf{w}^t - \mathbf{w}^\star\| \leq 1$$

- Since $\forall t : \|\mathbf{w}^t\| \leq 1 \wedge \|\mathbf{x}_i\| \leq 1$ then

$$G \leq 1 \quad D \leq 1$$

SGD ⇒ Perceptron

- "Ignore" rounds t such that $0 < y_t(\mathbf{w}^t \cdot \mathbf{x}^t) \leq \gamma$
- Loss bound becomes,

$$\gamma \sum_{t=1}^{T} \mathbb{1}\left[y_t(\mathbf{w}^t \cdot \mathbf{x}_t) \leq 0\right] \leq \sum_{t=1}^{T} [\gamma - y_t(\mathbf{w}^t \cdot \mathbf{x}_t)]_+$$

$$\leq \sum_{t=1}^{T} [\overbrace{\gamma - \underbrace{y_t(\mathbf{w}^\star \cdot \mathbf{x}_t)}_{\geq \gamma}}^{\leq 0}]_+ + \sqrt{T}$$

- If we saw only mistake-prone examples ⇒ $T = \#\text{mistakes}$

$$\gamma T \leq \sqrt{T} \quad \Rightarrow \quad T \leq \frac{1}{\gamma^2}$$

- SGD updates $\mathbf{w}^t$ on rounds when $y_t(\mathbf{w}^t \cdot \mathbf{x}^t)$ is small and is thus called the *aggressive* Perceptron

Logistic Regression

- Define the following estimate,

$$\mathbb{P}\left[Y = +1 | \mathbf{x}, \mathbf{w}\right] \stackrel{\text{def}}{=} \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{x})}$$

- We can write,

$$\begin{aligned} \mathbb{P}\left[Y = -1 | \mathbf{x}, \mathbf{w}\right] &= 1 - \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{x})} \\ &= \frac{1}{1 + \exp(\mathbf{w} \cdot \mathbf{x})} \end{aligned}$$

- Putting the two outcomes together we get,

$$\mathbb{P}\left[Y = y | \mathbf{x}, \mathbf{w}\right] \stackrel{\text{def}}{=} \frac{1}{1 + \exp(-y(\mathbf{w} \cdot \mathbf{x}))}$$

Logistic Regression

- Loss of wrong prediction,

$$-\log\left(\mathbb{P}\left[Y=-y_i|\mathbf{w},\mathbf{x}_i\right]\right)=-\log\left(1+e^{-y_i(\mathbf{w}\cdot\mathbf{x}_i)}\right)$$

- SGD iterate for sub-sample S'

$$\begin{aligned}\forall i\in S':\quad p_i &= \frac{1}{1+\exp(y_i(\mathbf{w}^t\cdot\mathbf{x}_i))}\\ \mathbf{g}^t &= -\sum_{i\in S'} p_i y_i \mathbf{x}_i\\ \mathbf{w}^{t+1} &= \mathbf{w}^t-\eta^t\mathbf{g}^t=\mathbf{w}^t+\eta^t\sum_{i\in S'} p_i\, y_i\, \mathbf{x}_i\end{aligned}$$