COS234:

Introduction To Machine Learning

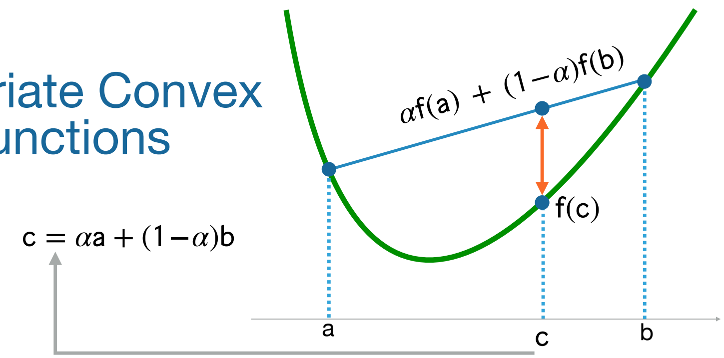
Prof. Yoram Singer

Princeton University

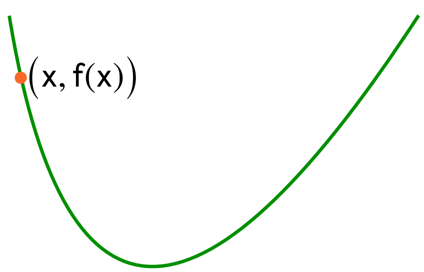
Topic: Gradient-Based Learning - Part II

$$\alpha \in [0,1]: \ f\big(\alpha a + (1-\alpha)b\big) \leq \alpha f(a) + (1-\alpha)f(b)$$

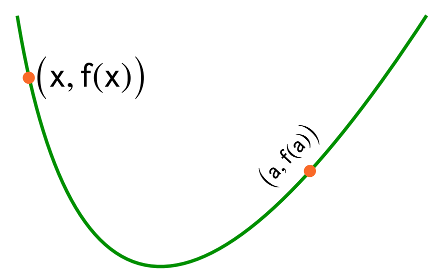
Univariate Convex Functions



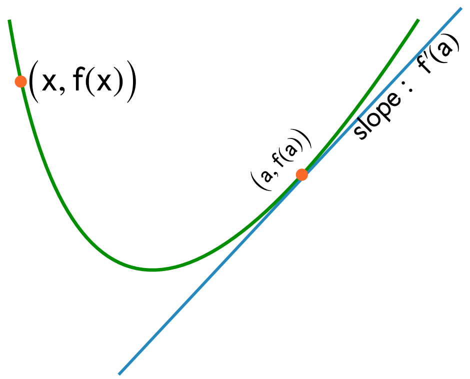
Univariate Convex Functions



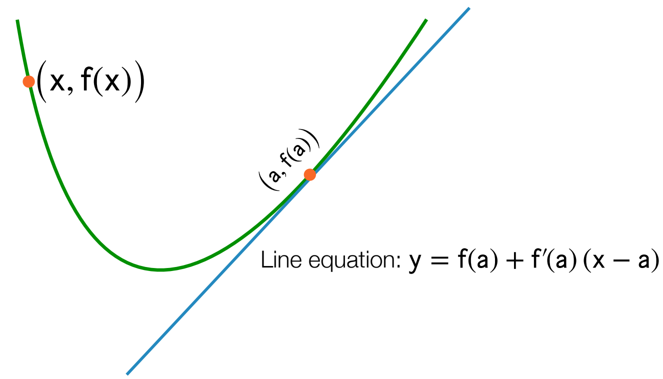
Univariate Convex Functions



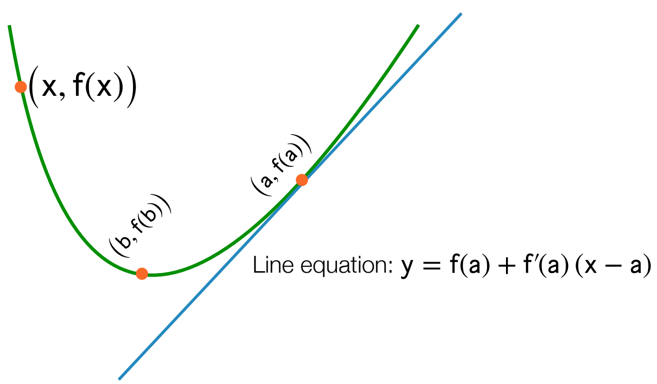
Univariate Convex Functions



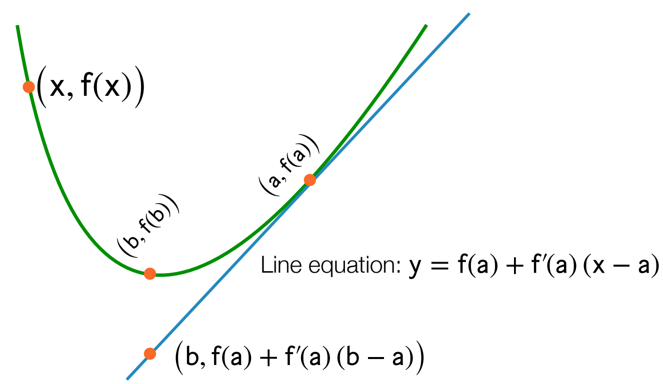
Univariate Convex Functions



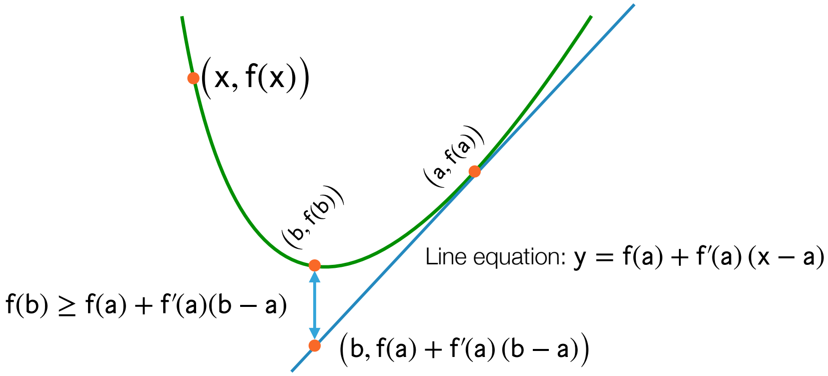
Univariate Convex Functions



Univariate Convex Functions



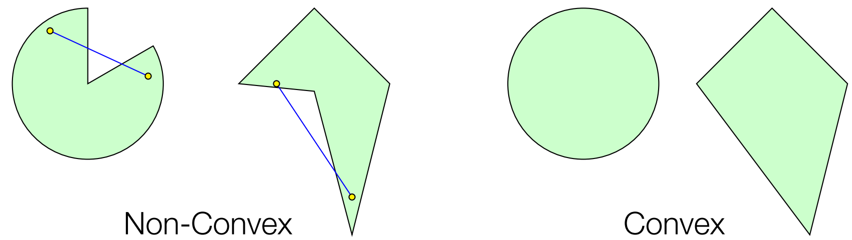
Univariate Convex Functions



Convex Sets

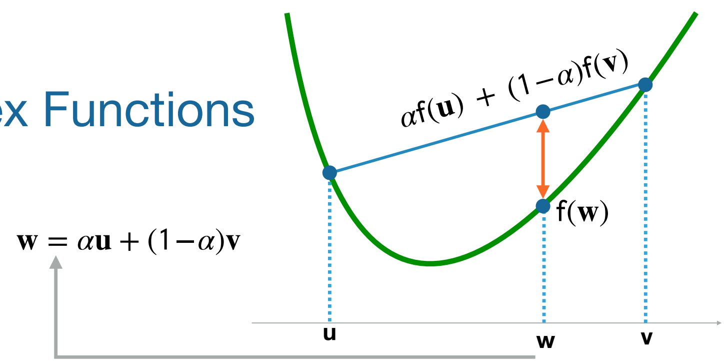
Set Ω is convex if for any $\mathbf{u}, \mathbf{v} \in \Omega$ line segment between $\mathbf{u}$ and $\mathbf{v}$ is in Ω:

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



$$\alpha \in [0, 1] : \; f\big(\alpha \mathbf{u} + (1 - \alpha)\mathbf{v}\big) \leq \alpha f(\mathbf{u}) + (1 - \alpha) f(\mathbf{v})$$

Convex Functions



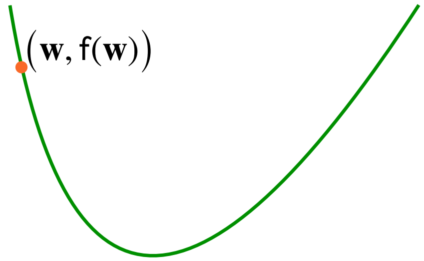
Gradients

$\mathscr{L}(\mathbf{w})$ is a function from $\mathbf{R}^{\mathbf{d}} \to \mathbf{R}_{+}$

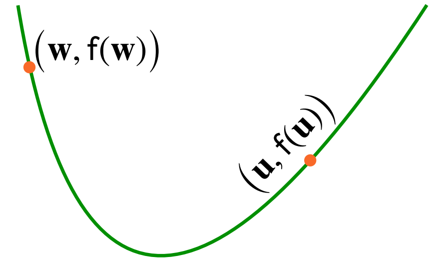
We need switch from *derivatives* to *gradients:*

$$\nabla \mathscr{L}(\mathbf{w}) \equiv \frac{\partial \mathscr{L}(\mathbf{w})}{\partial \mathbf{w}} = \left(\frac{\partial \mathscr{L}}{\partial w_1}, \frac{\partial \mathscr{L}}{\partial w_2}, \quad \ldots, \frac{\partial \mathscr{L}}{\partial w_d} \right)$$

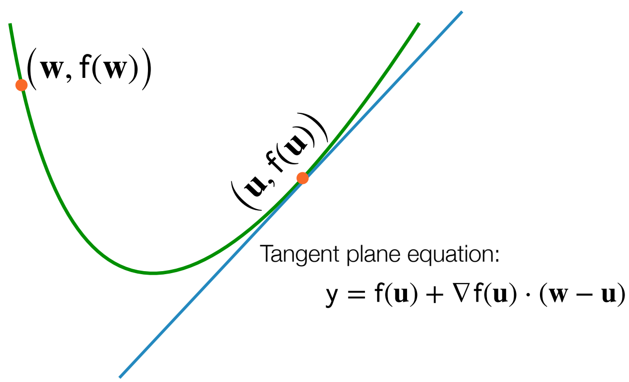
Multivariate Convex Functions



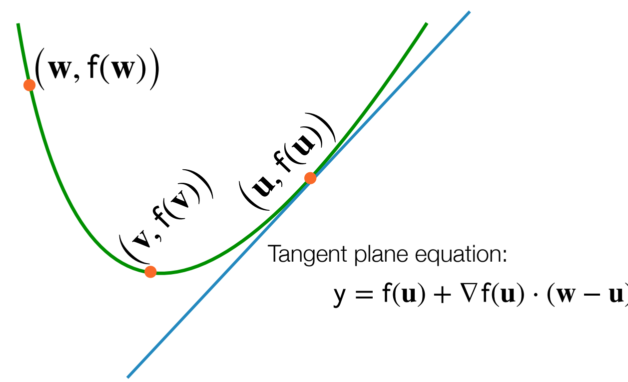
Multivariate Convex Functions



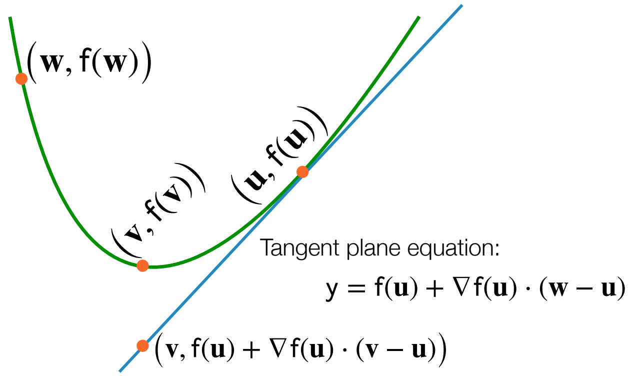
Multivariate Convex Functions



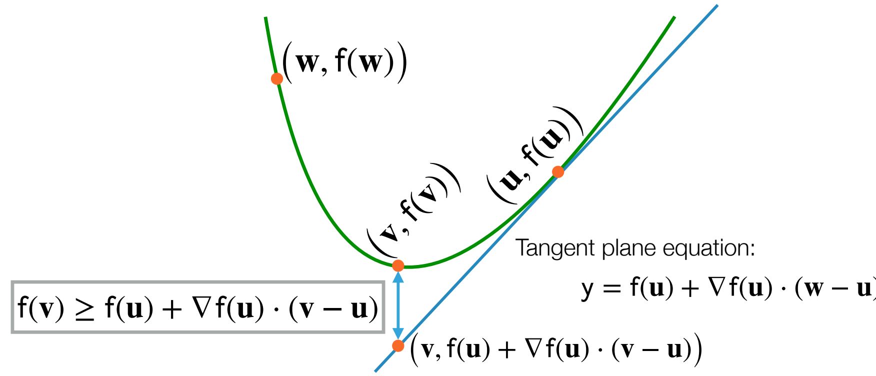
Multivariate Convex Functions



Multivariate Convex Functions



Multivariate Convex Functions



First Order Conditions

- Function $f : \mathbf{R}^d \rightarrow \mathbf{R}$ is convex iff:
  - For $\forall \alpha \in [0,1], \mathbf{u} \in \mathbf{R}^d, \mathbf{v} \in \mathbf{R}^d$:

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

  - $f(\mathbf{v}) \geq f(\mathbf{u}) + \nabla f(\mathbf{u}) \cdot (\mathbf{v} - \mathbf{u})$
  - Fix $\mathbf{u} \in \mathbf{R}^d, \mathbf{v} \in \mathbf{R}^d$ and define $h : \mathbf{R} \rightarrow \mathbf{R}$ as $h(t) = f(\mathbf{u} + t\mathbf{v})$ then $h(t)$ is a convex univariate function $\left(h''(t) \geq 0 \right)$

Second Order Conditions

Hessian of $f : \mathbf{R}^d \rightarrow \mathbf{R}$ is $d \times d$ matrix of second order derivatives $\nabla^2 f$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \, \partial w_j}$$

f is convex iff:

$$\forall \mathbf{u} : \ \mathbf{u}^T H \, \mathbf{u} \geq 0$$

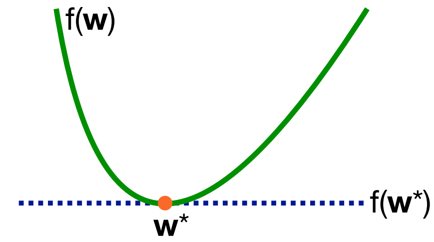
This means that the smallest Eigen value of H is non-negative

Second Order Conditions

- Function $f : \mathbf{R}^d \rightarrow \mathbf{R}$ is convex iff:
  - For $\forall \alpha \in [0, 1], \mathbf{u} \in \mathbf{R}^d, \mathbf{v} \in \mathbf{R}^d$:

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

  - $f(\mathbf{v}) \geq f(\mathbf{u}) + \nabla f(\mathbf{u}) \cdot (\mathbf{v} - \mathbf{u})$
  - Fix $\mathbf{u} \in \mathbf{R}^d, \mathbf{v} \in \mathbf{R}^d$ and define $h : \mathbf{R} \rightarrow \mathbf{R}$ as $h(t) = f(\mathbf{u} + t\mathbf{v})$



- $\mathbf{w}^*$ minimum (vector) of f:

$$f(\mathbf{w}^\star) \leq f(\mathbf{w}) \text{ for all } b \neq a^*$$

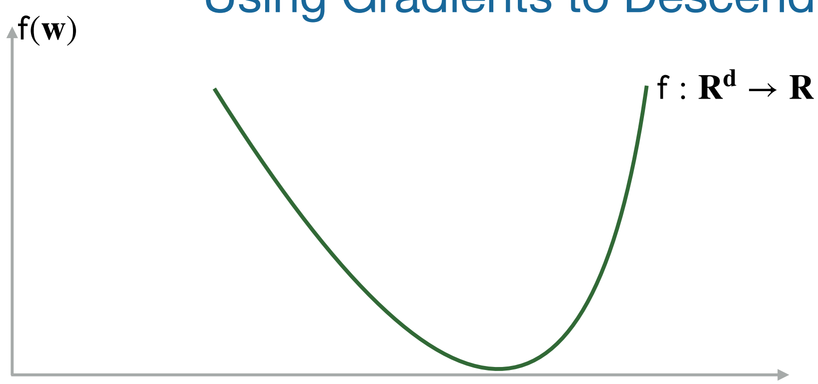
- Gradient of $f(\mathbf{w})$ at $\mathbf{w}^*$ is zero

$$\nabla f(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{w}^\star} = (0, 0, \ \ldots \ , 0) = \mathbf{0}$$

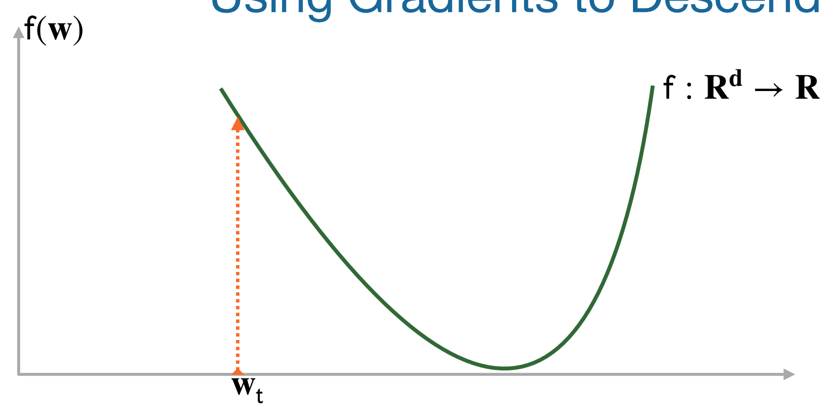
- Almost always no closed form solution for:

$$\nabla f(\mathbf{w}) = \mathbf{0}$$

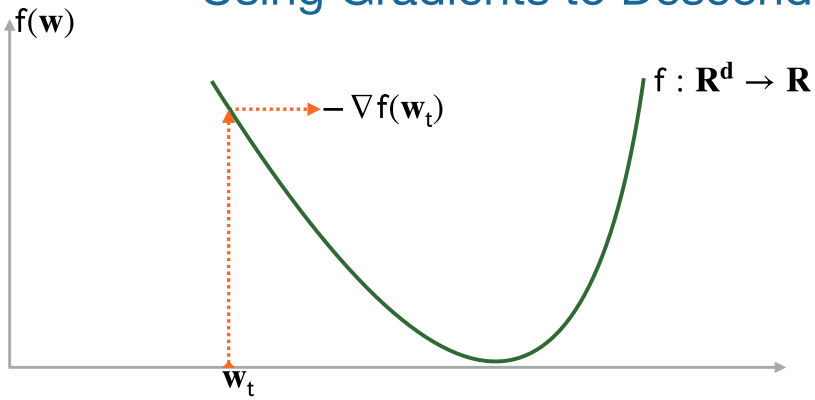
Using Gradients to Descend



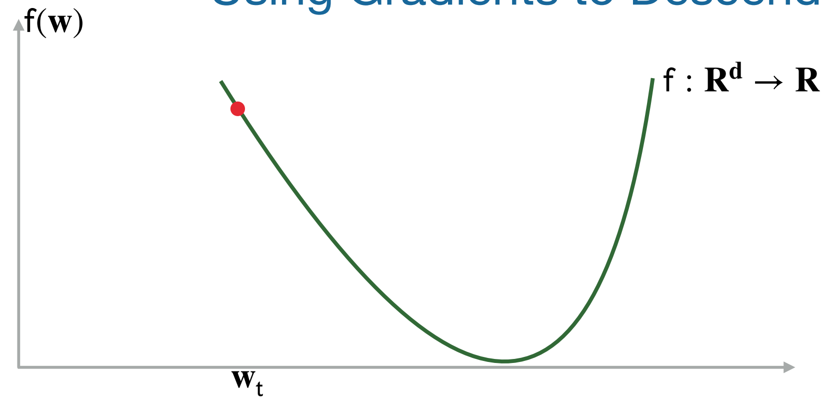
Using Gradients to Descend



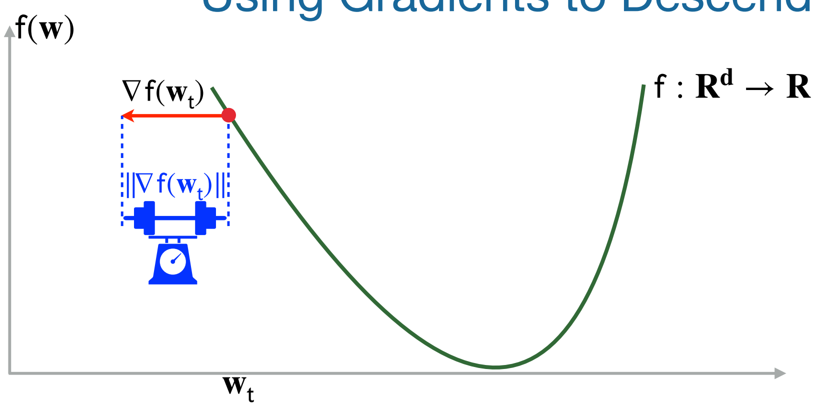
Using Gradients to Descend



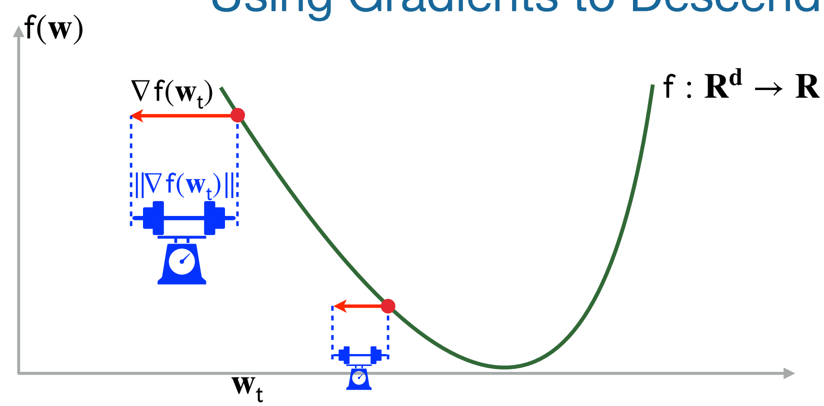
Using Gradients to Descend



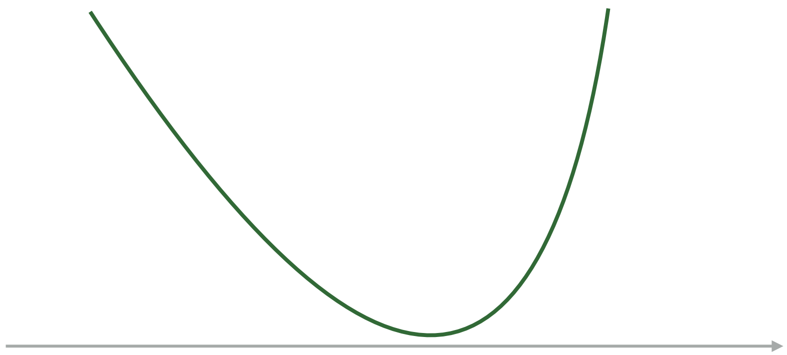
Using Gradients to Descend



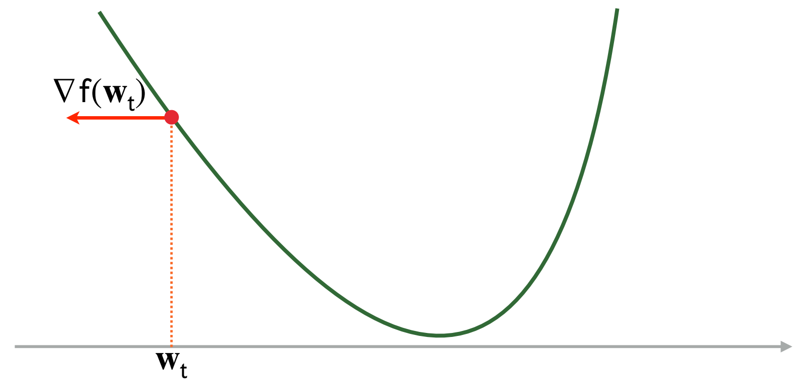
Using Gradients to Descend



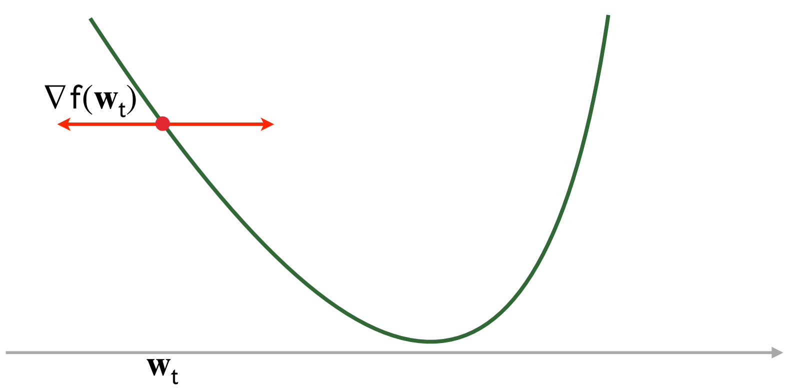
Gradient Descent



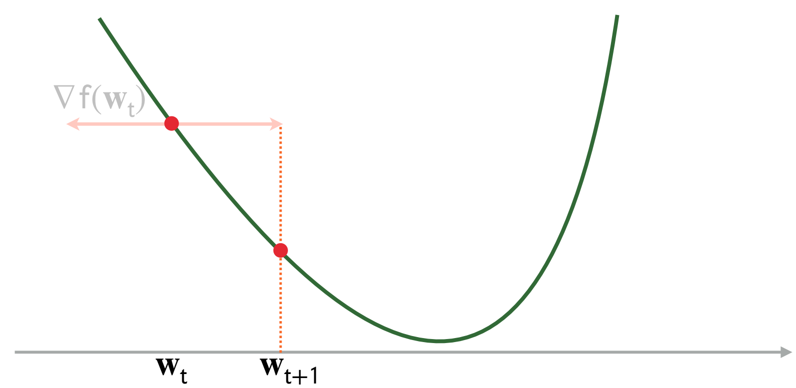
Gradient Descent



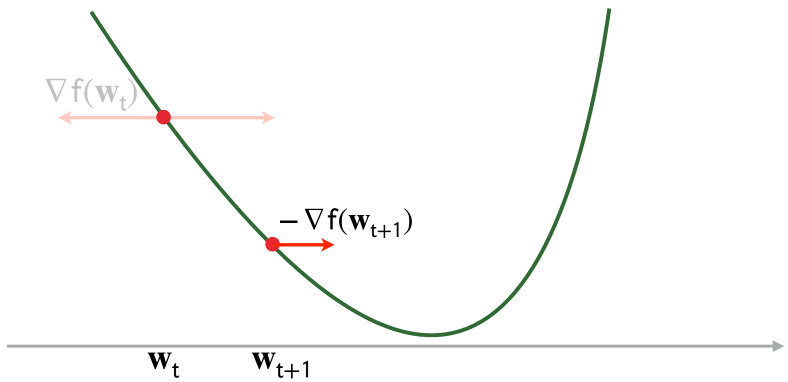
Gradient Descent



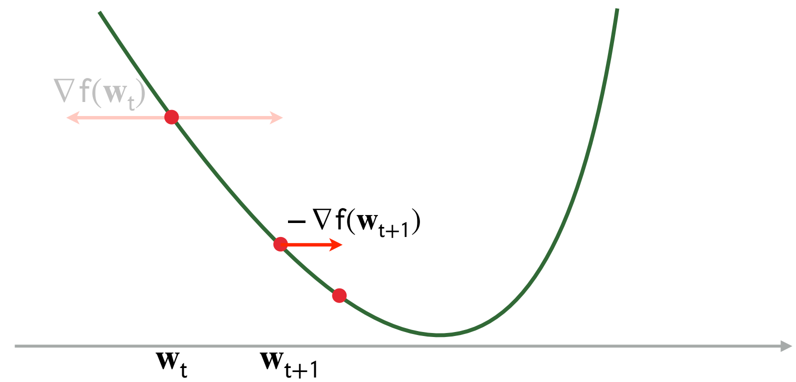
Gradient Descent



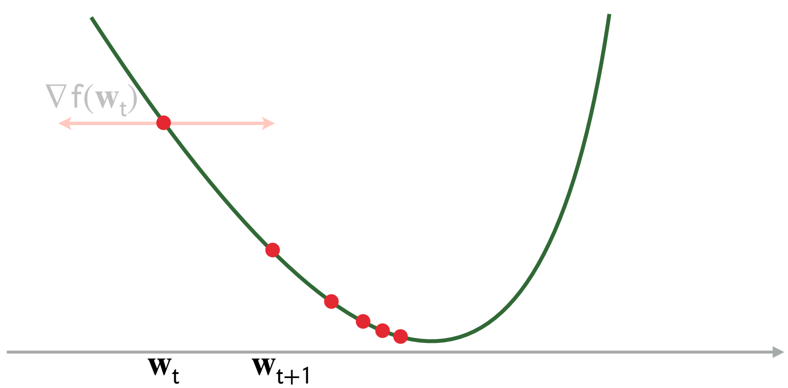
Gradient Descent



Gradient Descent



Gradient Descent



Gradient Descent Procedure

- Input: function $f : \mathbf{R} \to \mathbf{R}_+$
- Goal: find $\hat{\mathbf{w}}$ such that $f(\hat{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \epsilon$
- Choose initial value $\mathbf{w}_1$
- Loop:
  - $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)$
  - Until …

Gradient Descent Procedure

- Input: function $f : \mathbf{R} \to \mathbf{R}_+$
- Goal: find $\hat{\mathbf{w}}$ such that $f(\hat{\mathbf{w}}) - f(\mathbf{w}^\star) \le \epsilon$
- Choose initial value $\mathbf{w}_1$
- Loop:
  - $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)$
  - Until …

☆learning rate
☆step size

Gradient Descent

```
def derivative_descent(w0, gradient_func, eta):
    T, d = len(eta) + 1, len(w0)
    w = w0
    for t in range(1, T):
        w = w - eta[t-1] * gradient_func(w)
    return w
```


Learning Rate

- Crucial in many learning problems
- Fixed learning-rate can be used in restricted circumstances
- Self-tuning procedure of learning-rate exist, notably AdaGrad
- In many applications:
  - Linear decrease $\eta_t = \dfrac{\eta_0}{b + st}$ where $\eta_0 \in [0.1, 1]$ (typically)
  - Sub-linear decrease $\eta_t \sim \dfrac{\eta_0}{\sqrt{t}}$

Generalized Linear Models

Loss for linear predictors $\mathscr{L}(\mathbf{w}) = \dfrac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(\mathbf{w} \cdot \mathbf{x_i}))$

h: transfer (activation) function $h : \mathbf{R} \to \mathbf{R}$ and $\hat{y}_i = h(\mathbf{w} \cdot \mathbf{x_i})$

$\ell(y, \hat{y})$ binary function from to $\mathbf{R} \times \mathbf{R}$ to $\mathbf{R}_+$

We can find $\mathbf{w}$ which (approximately) minimizes $\mathscr{L}(\mathbf{w})$ using GD and computing the gradient using chain rule

GLM: Examples

Linear regression:

$y \in \mathbf{R} \quad h(z) = z \quad \ell(y, \hat{y}) = (y - \hat{y})^2$

Classification with hinge-loss:

$y \in \{-1, +1\} \quad h(z) = z \quad \ell(y, \hat{y}) = [1 - y\hat{y}]_+$ where $[z]_+ = \max\{0, z\}$

Logistic regression:

$y \in \{0, 1\} \quad h(z) = \dfrac{1}{1 + e^{-z}} \quad \ell(y, \hat{y}) = -y\log(\hat{y}) - (1 - y)\log(1 - \hat{y})$

Gradients for GLM

$$\nabla \mathscr{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{w}}\, \ell(y_i, h(\mathbf{w} \cdot \mathbf{x_i}))$$

Given y_i derivative w.r.t $\hat{y}_i$ is $\ell'(y_i, \hat{y}_i) = \dfrac{d\,\ell(y_i, \hat{y}_i)}{d\hat{y}_i}$

Define $z_i = \mathbf{w} \cdot \mathbf{x}_i$ then derivative of h w.r.t. z_i is $h'(z_i)$

Finally $\nabla_{\mathbf{w}}\, z_i = \mathbf{x}_i$

Thus $\nabla \mathscr{L}(\mathbf{w}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \ell'(y_i, \hat{y}_i)\, h'(z_i)\, \mathbf{x_i}$

Gradients for GLM

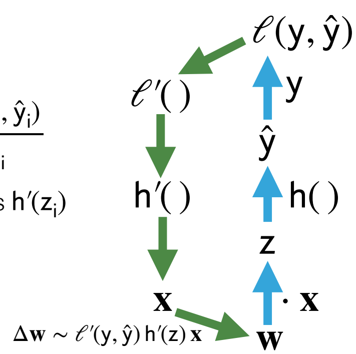
$$\nabla \mathscr{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{w}}\, \ell(y_i, h(\mathbf{w} \cdot \mathbf{x_i}))$$

Given y_i derivative w.r.t $\hat{y}_i$ is $\ell'(y_i, \hat{y}_i) = \dfrac{d\,\ell(y_i, \hat{y}_i)}{d\hat{y}_i}$

Define $z_i = \mathbf{w} \cdot \mathbf{x}_i$ then derivative of h w.r.t. z_i is $h'(z_i)$

Finally $\nabla_{\mathbf{w}}\, z_i = \mathbf{x}_i$

Thus $\nabla \mathscr{L}(\mathbf{w}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \ell'(y_i, \hat{y}_i)\, h'(z_i)\, \mathbf{x_i}$



Gradient for Logistic Regression

Gradient for Logistic Regression

$$y \in \{0,1\} \quad h(z) = \frac{1}{1+e^{-z}} \quad \ell(y,\hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

Gradient for Logistic Regression

$$y \in \{0,1\} \quad h(z) = \frac{1}{1+e^{-z}} \quad \ell(y,\hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

$$\ell'(y,\hat{y}) = \frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

Gradient for Logistic Regression

$$y \in \{0,1\} \quad h(z) = \frac{1}{1+e^{-z}} \quad \ell(y,\hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

$$\ell'(y,\hat{y}) = \frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y}})$$

$$h'(z) = h(z)\,(1-h(z)) = \hat{y}\,(1-\hat{y})$$

Gradient for Logistic Regression

$$y \in \{0,1\} \quad h(z) = \frac{1}{1+e^{-z}} \quad \ell(y,\hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

$$\ell'(y,\hat{y}) = \frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

$$h'(z) = h(z)\,(1-h(z)) = \hat{y}\,(1-\hat{y})$$

$$\Rightarrow \quad \ell'(y,\hat{y})\,h'(z) = \hat{y} - y$$

Gradient for Logistic Regression

$$y \in \{0,1\} \quad h(z) = \frac{1}{1+e^{-z}} \quad \ell(y,\hat{y}) = -\,y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

$$\ell'(y,\hat{y}) = \frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

$$h'(z) = h(z)\,(1-h(z)) = \hat{y}\,(1-\hat{y})$$

$$\ell'(y,\hat{y})\,h'(z) = \hat{y} - y$$

$$\nabla \mathscr{L}(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)\,\mathbf{x}_i$$

Gradient for Logistic Regression

$$y \in \{0,1\} \quad h(z) = \frac{1}{1+e^{-z}} \quad \ell(y,\hat{y}) = -\,y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

$$\ell'(y,\hat{y}) = \frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

$$h'(z) = h(z)\,(1-h(z)) = \hat{y}\,(1-\hat{y})$$

$$\ell'(y,\hat{y})\,h'(z) = \hat{y} - y$$

$$\nabla \mathscr{L}(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n} \boxed{(\hat{y}_i - y_i)}\,\mathbf{x}_i$$

Interpretation?

Properties of GD

Assume that $\|\nabla \mathscr{L}(\mathbf{w})\| \le \gamma$

This typically amounts to $\forall i:\ \|\mathbf{x}_i\| \le c\gamma$

Assume that $\forall t:\ \|\mathbf{w}_t\| \le r$ and $\|\mathbf{w}^\star\| \le r$

Recall that from convexity $f(\mathbf{v}) \ge f(\mathbf{u}) + \nabla f(\mathbf{u}) \cdot (\mathbf{v} - \mathbf{u})$

Using $\mathbf{v} = \mathbf{w}^\star$ and $\mathbf{u} = \mathbf{w}_t$ we get

$$\mathscr{L}(\mathbf{w}_t) - \mathscr{L}(\mathbf{w}^\star) \le \nabla\mathscr{L}(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^\star)$$

Properties of GD (cont)

Going forward we abbreviate $\mathbf{g}_t = \nabla \mathscr{L}(\mathbf{w}_t)$ and thus we can write

$$\mathscr{L}(\mathbf{w}_t) - \mathscr{L}(\mathbf{w}^\star) \le \mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^\star)$$

Define progress towards (unknown) optimum as $\Delta_t = \|\mathbf{w}_t - \mathbf{w}^\star\|^2$

From GD update & boundedness of gradients:

$$\Delta_{t+1} \le \Delta_t - 2\eta\mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^\star) + \eta^2\gamma^2$$

Rearranging terms and dividing by 2η gives

$$\mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^\star) \le \frac{1}{2\eta}(\Delta_t - \Delta_{t+1}) + \frac{\eta\gamma^2}{2}$$

Properties of GD (cont)

Going forward we abbreviate $\mathbf{g}_t = \nabla \mathscr{L}(\mathbf{w}_t)$ and thus we can write

$$\mathscr{L}(\mathbf{w}_t) - \mathscr{L}(\mathbf{w}^\star) \le \mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^\star) \le \frac{1}{2\eta}(\Delta_t - \Delta_{t+1}) + \frac{\eta\gamma^2}{2}$$

Define progress towards (unknown) optimum as $\Delta_t = \|\mathbf{w}_t - \mathbf{w}^\star\|^2$

From GD update & boundedness of gradients:

$$\Delta_{t+1} \le \Delta_t - 2\eta \mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^\star) + \eta^2\gamma^2$$

Rearranging terms and dividing by 2η gives

$$\mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^\star) \le \frac{1}{2\eta}(\Delta_t - \Delta_{t+1}) + \frac{\eta\gamma^2}{2}$$

We established that $\mathscr{L}(\mathbf{w}_t) - \mathscr{L}(\mathbf{w}^\star) \le \frac{1}{2\eta}(\Delta_t - \Delta_{t+1}) + \frac{\eta\gamma^2}{2}$

Taking the average from t=1 through T

$$\frac{1}{T}\sum_{t=1}^{T} \mathscr{L}(\mathbf{w}_t) - \mathscr{L}(\mathbf{w}^\star) \le \frac{1}{2\eta T}(\Delta_1 - \Delta_{T+1}) + \frac{\eta\gamma^2}{2} \le \frac{r^2}{\eta T} + \frac{\eta\gamma^2}{2}$$

Assume $\forall t : \mathscr{L}(\mathbf{w}_t) \le \mathscr{L}(\mathbf{w}_{t-1})$ then $\mathscr{L}(\mathbf{w}_T) - \mathscr{L}(\mathbf{w}^\star) \le \frac{r^2}{\eta T} + \frac{\eta\gamma^2}{2}$

Choosing $\eta = r/(\gamma\, T^{1/2})$ gives $\mathscr{L}(\mathbf{w}_T) - \mathscr{L}(\mathbf{w}^\star) \le \frac{2r\gamma}{\sqrt{T}}$

Convergence of GD

Assumptions:

$\|\nabla \mathscr{L}(\mathbf{w})\| \le \gamma$, $\forall t : \|\mathbf{w}_t\| \le r$, $\|\mathbf{w}^\star\| \le r$

$\mathscr{L}(\mathbf{w}_t)$ is monotonically decreasing

Using learning rate of $\eta = r/(\gamma\sqrt{t})$

Then $\mathscr{L}(\mathbf{w}_t)$ converges t to $\mathscr{L}(\mathbf{w}^\star)$ at a rate of $O\left(\gamma\, r\, t^{-1/2}\right)$

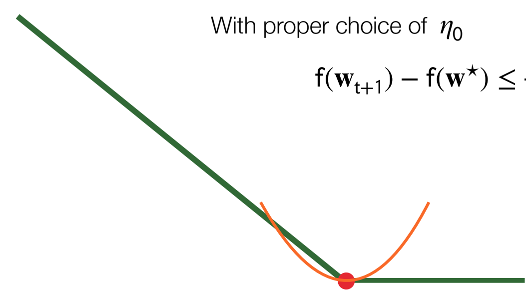
Summary: GD for Convex Losses

Learning rate goes to zero $\eta_0\, t^{-1/2}$

With proper choice of η_0

$$f(\mathbf{w}_{t+1}) - f(\mathbf{w}^\star) \le \frac{\kappa}{\sqrt{t}}$$

problem dependent



Next

When data is prevalent:

Stochastic Gradient Descent (SGD)

Solving multiclass problems with SGD

Beyond GLIM: deep learning