Non-convex Optimization for Machine Learning

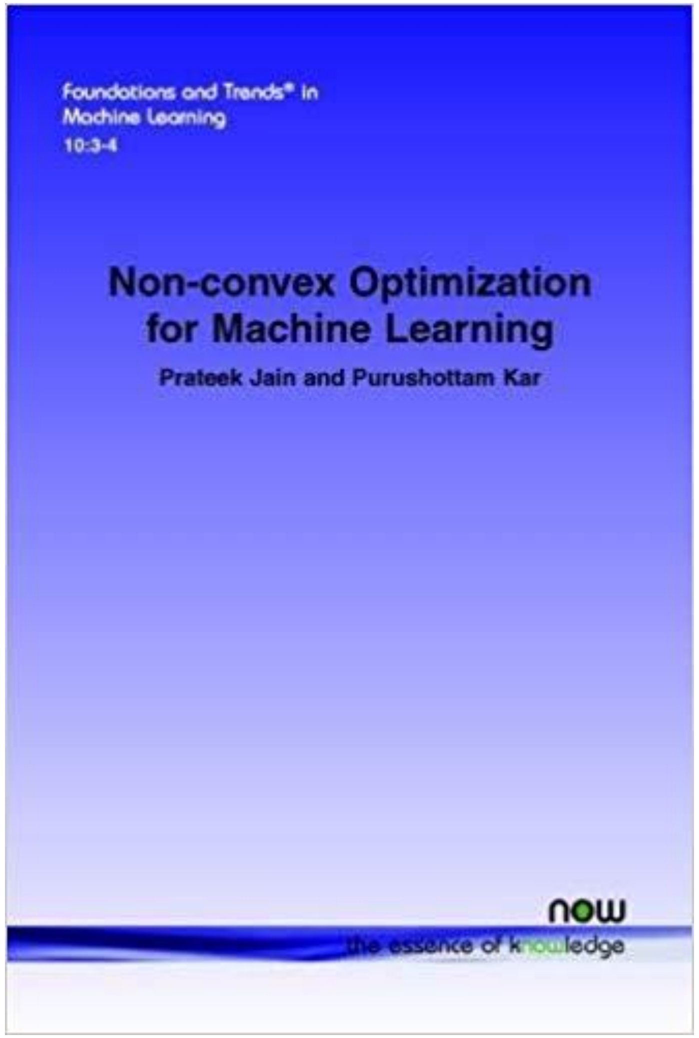
Prateek Jain

Microsoft Research India

Outline

- Optimization for Machine Learning
- Non-convex Optimization
- Convergence to Stationary Points
  - First order stationary points
  - Second order stationary points
- Non-convex Optimization in ML
  - Neural Networks
  - Learning with Structure
    - Alternating Minimization
    - Projected Gradient Descent

Relevant Monograph (Shameless Ad)



Optimization in ML

Supervised Learning

- Given points (x_i, y_i)
- Prediction function: $\widehat{y_i} = \phi(x_i, w)$
- Minimize loss: $\min_w \sum_i \ell(\phi(x_i, w), y_i)$

Unsupervised Learning

Given points $(x_1, x_2 \ldots x_N)$

Find cluster center or train GANs

Represent $\widehat{x_i} = \phi(x_i, w)$

Minimize loss: $\min_w \sum_i \ell(\phi(x_i, w), x_i)$

Optimization Problems

- Unconstrained optimization
$$\min_{w \in R^d} f(w)$$

- Deep networks
- Regression
- Gradient Boosted Decision Trees

- Constrained optimization
$$\min_{w} f(w) \;\; s.t. w \in C$$

- Support Vector Machines
- Sparse regression
- Recommendation system
- …

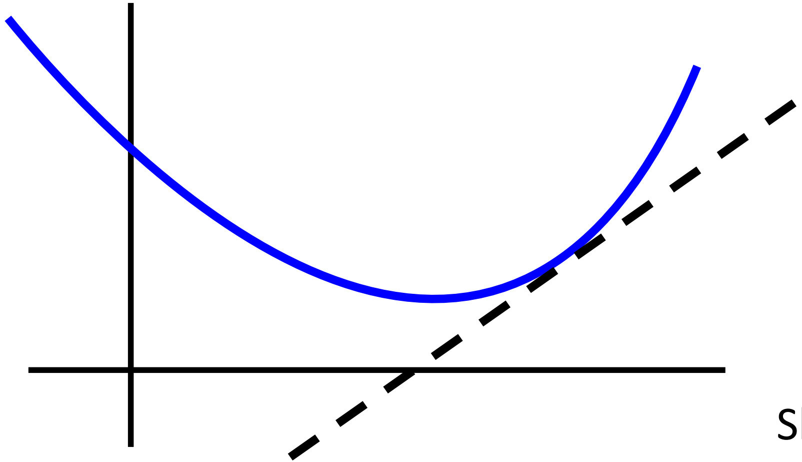
Convex Optimization

$$\min_{w} f(w)$$
$$s.t.\ \ w \in C$$

$$f : \mathbb{R}^d \to \mathbb{R}$$

Convex function

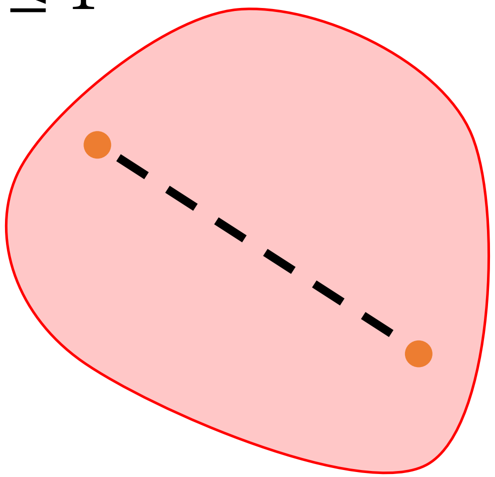
$f(\lambda w_1 + (1-\lambda) w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2),$
$0 \leq \lambda \leq 1$



$$C \subseteq \mathbb{R}^d$$

Convex set

$\forall w_1, w_2 \in C, \lambda w_1 + (1-\lambda) w_2 \in C$
$0 \leq \lambda \leq 1$



Slide credit: Purushottam Kar

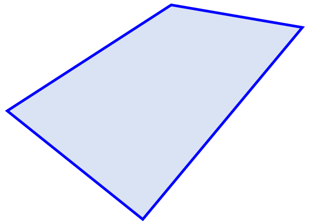
Examples

Linear Programming

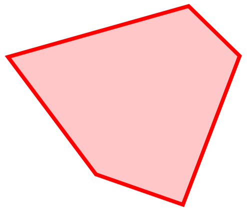
$$\min_{\mathbf{x}\in\mathbb{R}^d} \ \mathbf{a}^\top \mathbf{x}$$

$$s.t. \ \mathbf{b}_i^\top \mathbf{x} \leq c_i$$

f



$\mathcal{C}$

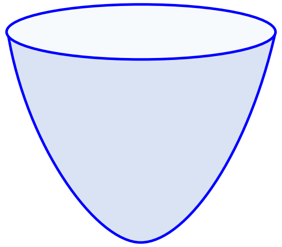


Quadratic Programming

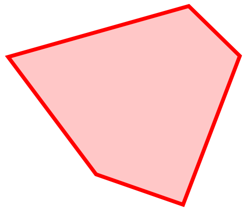
$$\min_{\mathbf{x}\in\mathbb{R}^d} \ \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{a}^\top \mathbf{x}$$

$$s.t. \ \mathbf{b}_i^\top \mathbf{x} \leq c_i$$

f



$\mathcal{C}$

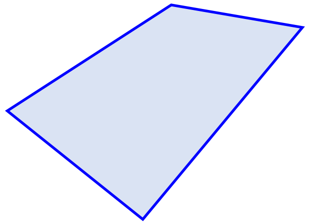


Semidefinite Programming

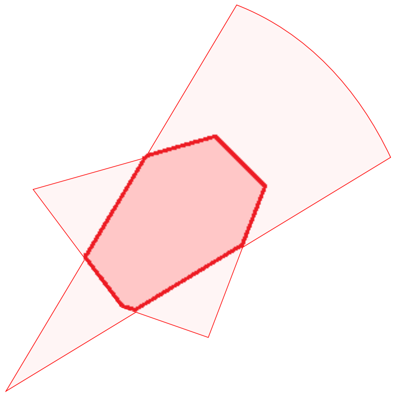
$$\min_{\mathbf{X}\succeq\mathbf{0}} \ \mathbf{A}^\top \mathbf{X}$$

$$s.t. \ \mathbf{B}_i^\top \mathbf{X} \leq c_i$$

f



$\mathcal{C}$

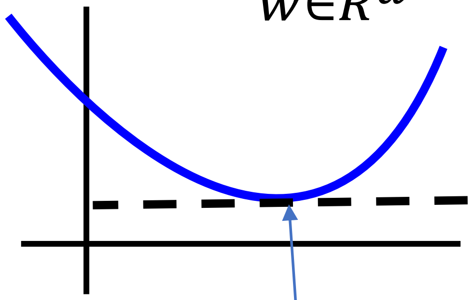


Slide credit: Purushottam Kar

Convex Optimization

- Unconstrained optimization

$$\min_{w \in R^d} f(w)$$



Optima: just ensure

$$\nabla_w f(w) = 0$$

- Constrained optimization

$$\min_{w} f(w) \;\; s.t. \, w \in C$$

Optima: KKT conditions

In this talk, lets assume f is L −smooth => f is differentiable

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} ||x - y||^2$$

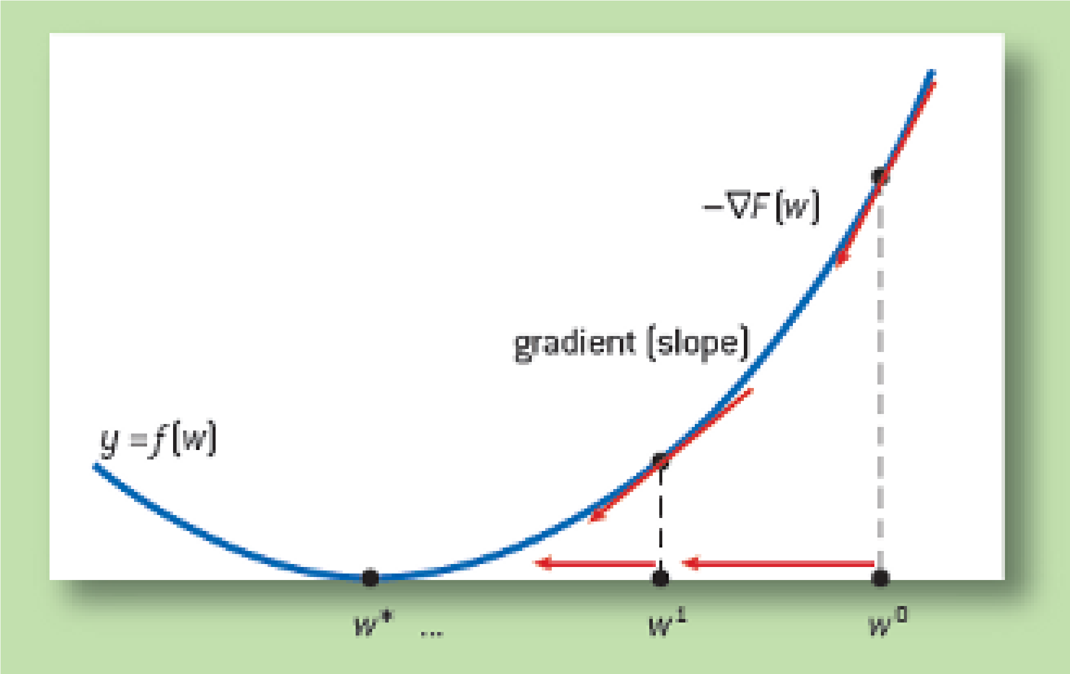
$$\text{OR,} \qquad ||\nabla f(x) - \nabla f(y)|| \leq L ||x - y||$$

Gradient Descent Methods

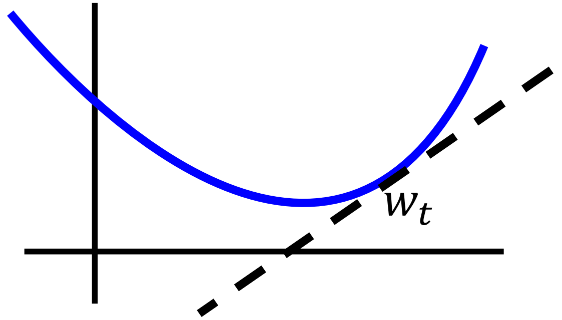
- Projected gradient descent method:

- For t=1, 2, … (until convergence)
  - $w_{t+1} = P_C(w_t - \eta\nabla f(w_t))$

- η: step-size



Convergence Proof



(a) $w_{t+1} = w_t - \eta \nabla f(w_t)$

(b) $\eta < \frac{1}{L}$

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2}||w_{t+1} - w_t||^2$$

$$f(w_{t+1}) \leq f(w_t) - \left(1 - \frac{L\eta}{2}\right)\eta||\nabla f(w_t)||^2 \leq f(w_t) - \frac{\eta}{2}||\nabla f(w_t)||^2$$

$$f(w_{t+1}) \leq f(w_*) + \langle \nabla f(w_t), w_t - w_* \rangle - \frac{1}{2\eta}||w_{t+1} - w_t||^2$$

Convexity

$$f(w_T) \leq f(w_{t+1}) \leq f(w_*) + \frac{1}{2\eta}\left(||w_t - w_*||^2 - ||w_{t+1} - w_*||^2\right)$$

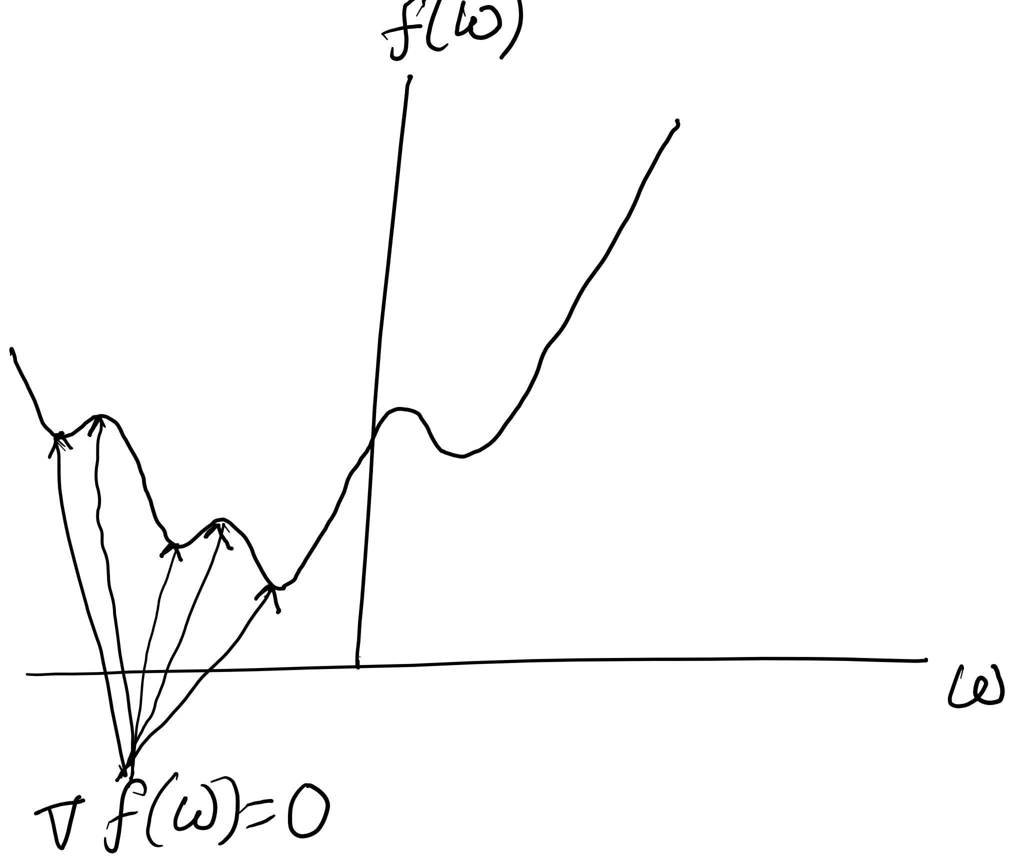
$$f(w_T) \leq f(w_*) + \frac{1}{T \cdot 2\eta}||w_0 - w_*||^2 \Rightarrow f(w_T) \leq f(w_*) + \epsilon$$

$$T = O\left(\frac{L \cdot ||w_0 - w_*||^2}{\epsilon}\right)$$

Non-convexity?

$$\min_{w \in R^d} f(w)$$

- Critical points: $\nabla f(w) = 0$
- But: $\nabla \mathrm{f}(\mathrm{w}) = 0 \nRightarrow$ Optimality



Local Optima

- $f(w) \leq f(w'), \forall ||w - w'|| \leq \epsilon$

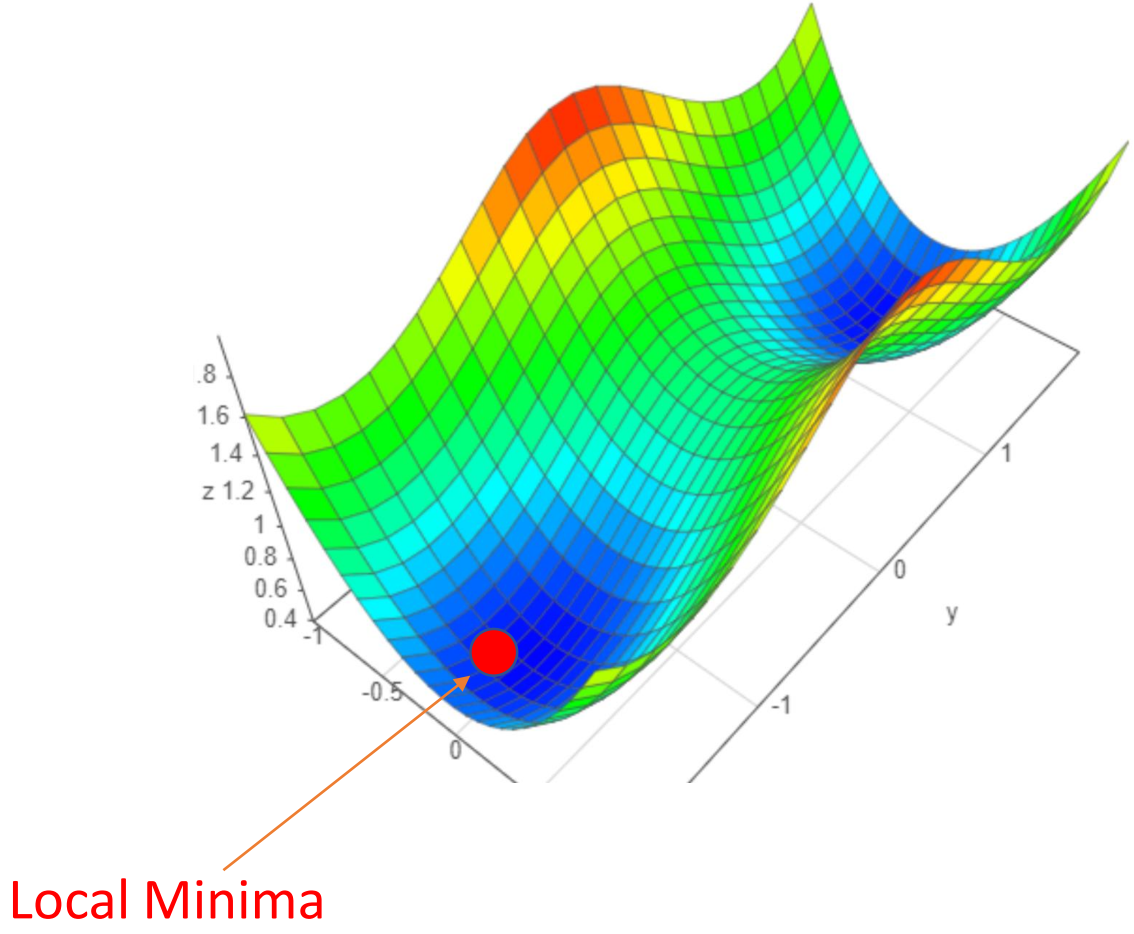
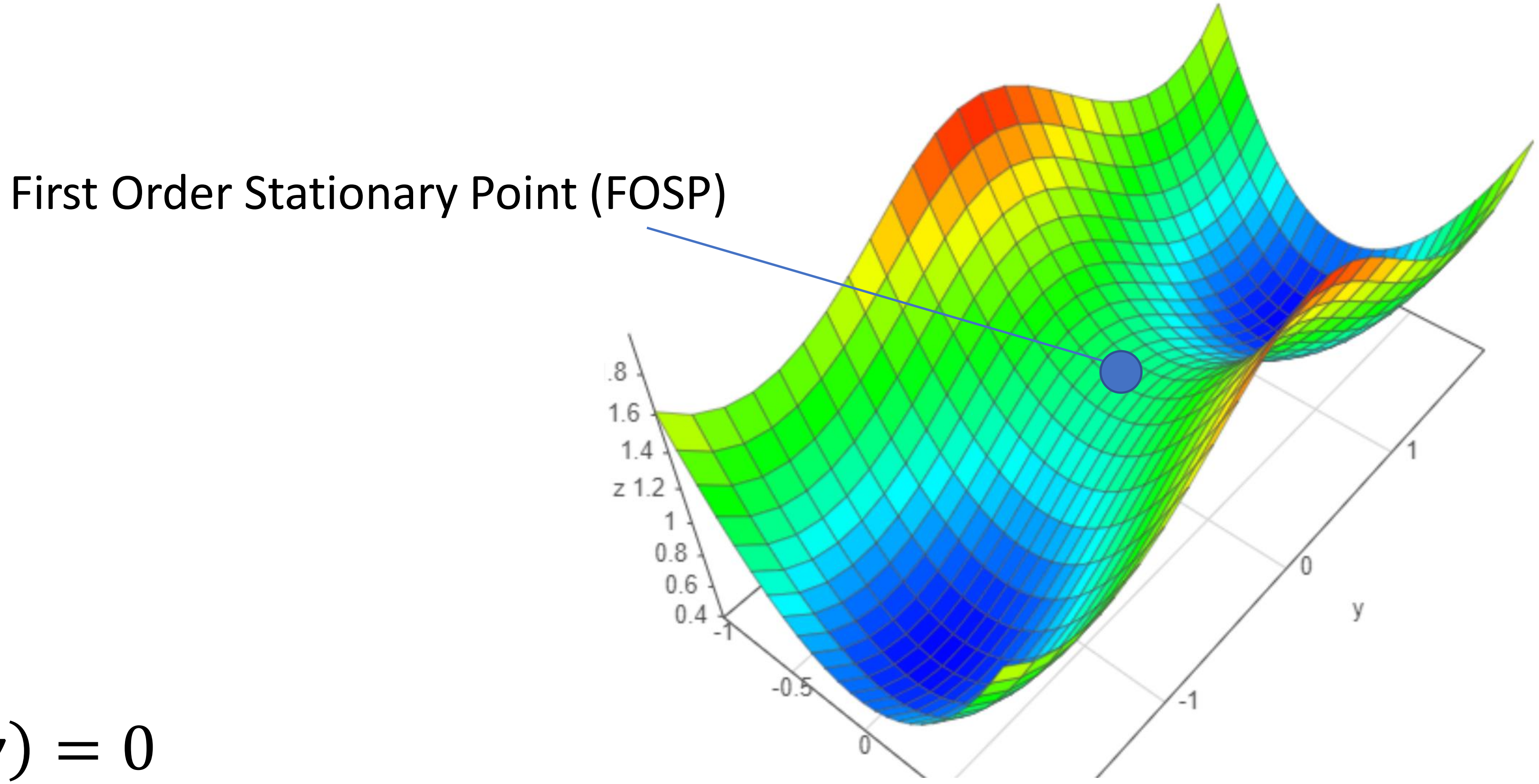


image credit: academo.org

First Order Stationary Points

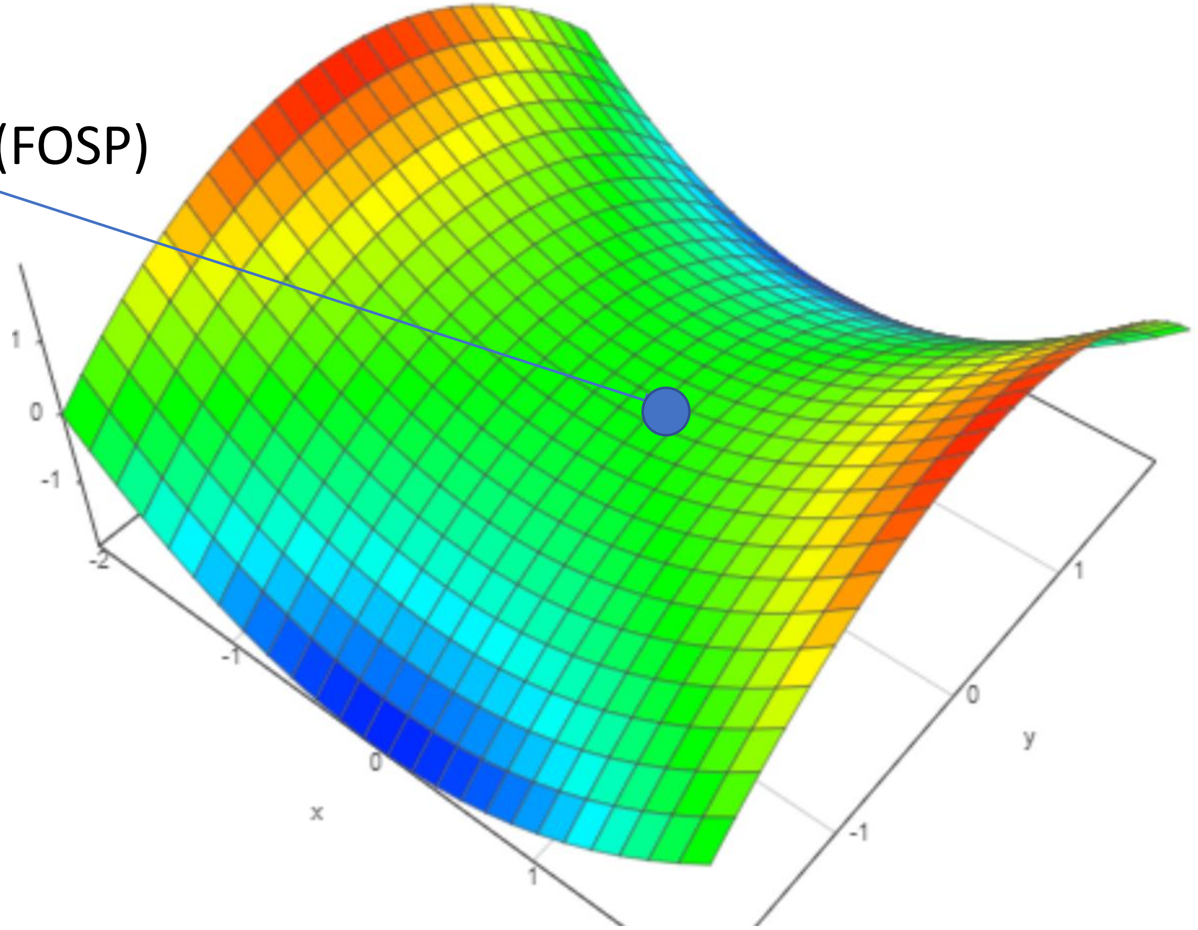


- Defined by: $\nabla f(w) = 0$

- But $\nabla^2 f(w)$ need not be positive semi-definite

image credit: academo.org

First Order Stationary Points



- E.g., $f(w) = 0.5(w_1^2 - w_2^2)$
- $\nabla f(w) = \begin{bmatrix} w_1 \\ -w_2 \end{bmatrix}$
- $\nabla f(0) = 0$
- But, $\nabla^2 f(w) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \Rightarrow indefinite$

- $f\left(\left[\frac{\epsilon}{2}, \epsilon\right]\right) = -\frac{3}{8}\,\epsilon^2 \Rightarrow f([0,0])$ is not a local minima

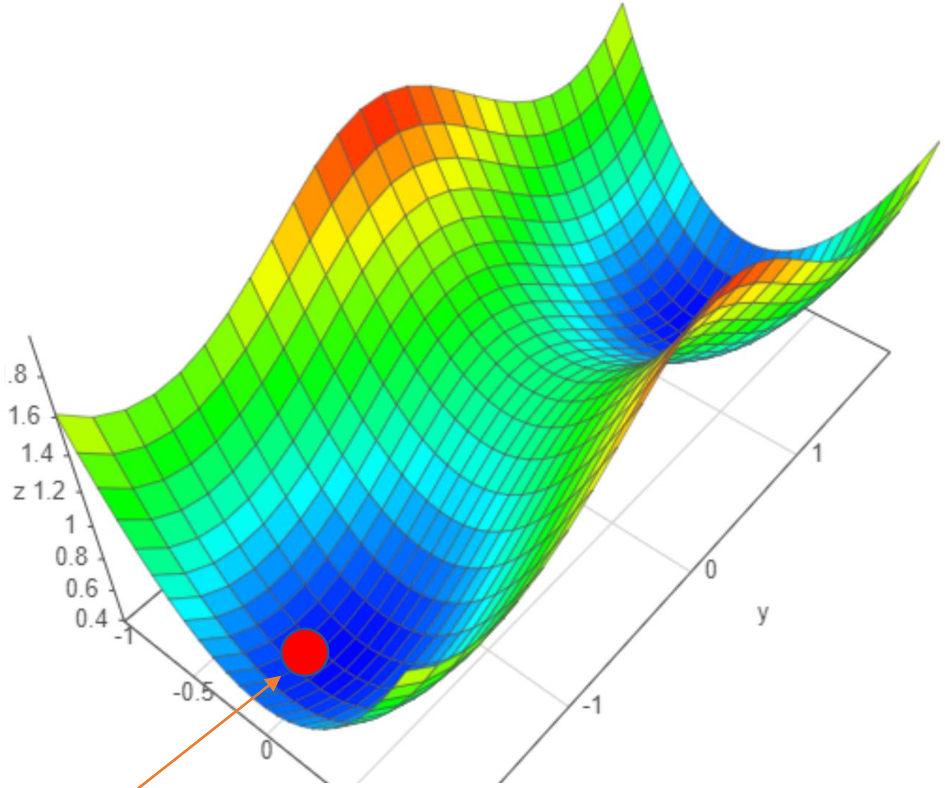
image credit: academo.org

Second Order Stationary Points

Second Order Stationary Point (SOSP) if:

- $\nabla f(w) = 0$
- $\nabla^2 f(w) \succcurlyeq 0$

Does it imply local optimality?

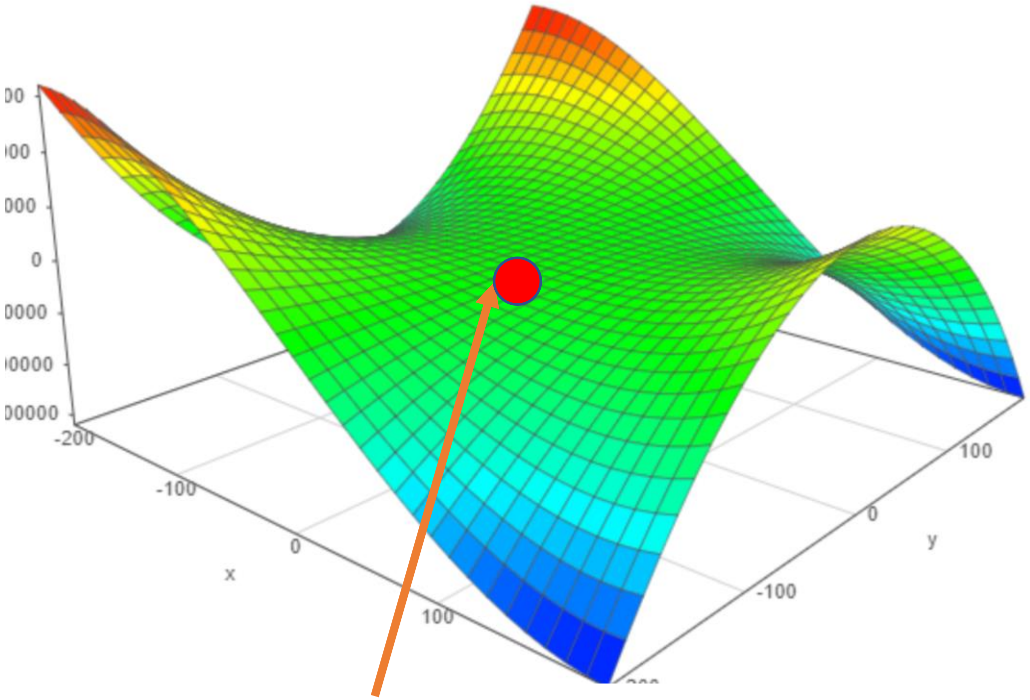


Second Order Stationary Point (SOSP)

image credit: academo.org

Second Order Stationary Points

- $f(w) = \frac{1}{3}(w_1^3 - 3\, w_1 w_2^2)$
- $\nabla f(w) = \begin{bmatrix} (w_1^2 - w_2^2) \\ -2\, w_1 w_2 \end{bmatrix}$
- $\nabla^2 f(w) = \begin{bmatrix} 2w_1 & -2w_2 \\ -2w_2 & -2w_1 \end{bmatrix}$
- $\nabla f(0) = 0, \nabla^2 f(0) = 0 \Rightarrow 0\ is\ SOSP$
- $f([\epsilon, \epsilon]) = -\frac{2}{3}\epsilon^3 < f(0)$



Second Order Stationary Point (SOSP)

image credit: academo.org

Stationarity and local optima

- w is local optima implies: $f(w) \leq f(w'),\ \forall ||w - w'|| \leq \epsilon$

- w is FOSP implies:
$$f(w) \leq f(w') + O(||w - w||^2)$$
- w is SOSP implies:
$$f(w) \leq f(w') + O(||w - w'||^3)$$
- w is p-th order SP implies:
$$f(w) \leq f(w') + O(||w - w'||^{p+1})$$

- That is, local optima: $p = \infty$

Computability? $f(w) \le f(w') + O(||w - w'||^{p+1})$

First Order Stationary Point	✓	
Second Order Stationary Point	✓	
Third Order Stationary Point	✓	
$p \ge 4$ Stationary Point	✗	NP-Hard
Local Optima	✗	NP-Hard

Anandkumar and Ge-2016

Does Gradient Descent Work for Local Optimality?

- Yes!
- In fact, with high probability converges to a "local minimizer"
  - If initialized randomly!!!
- But no rates known ☹
  - NP-hard in general!!
  - Big open problem ☺

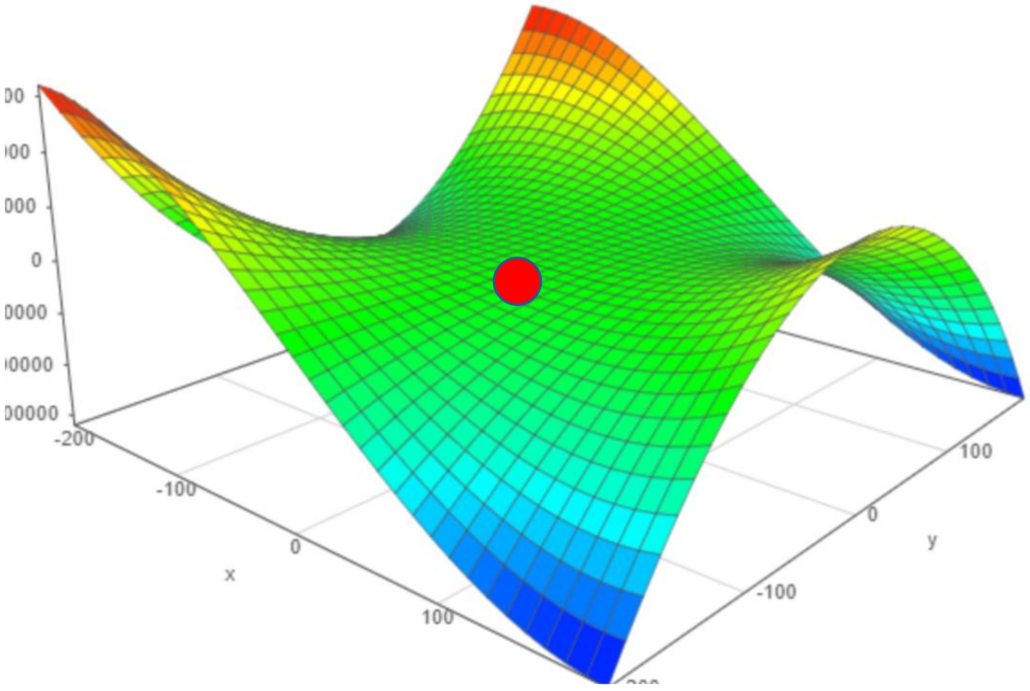
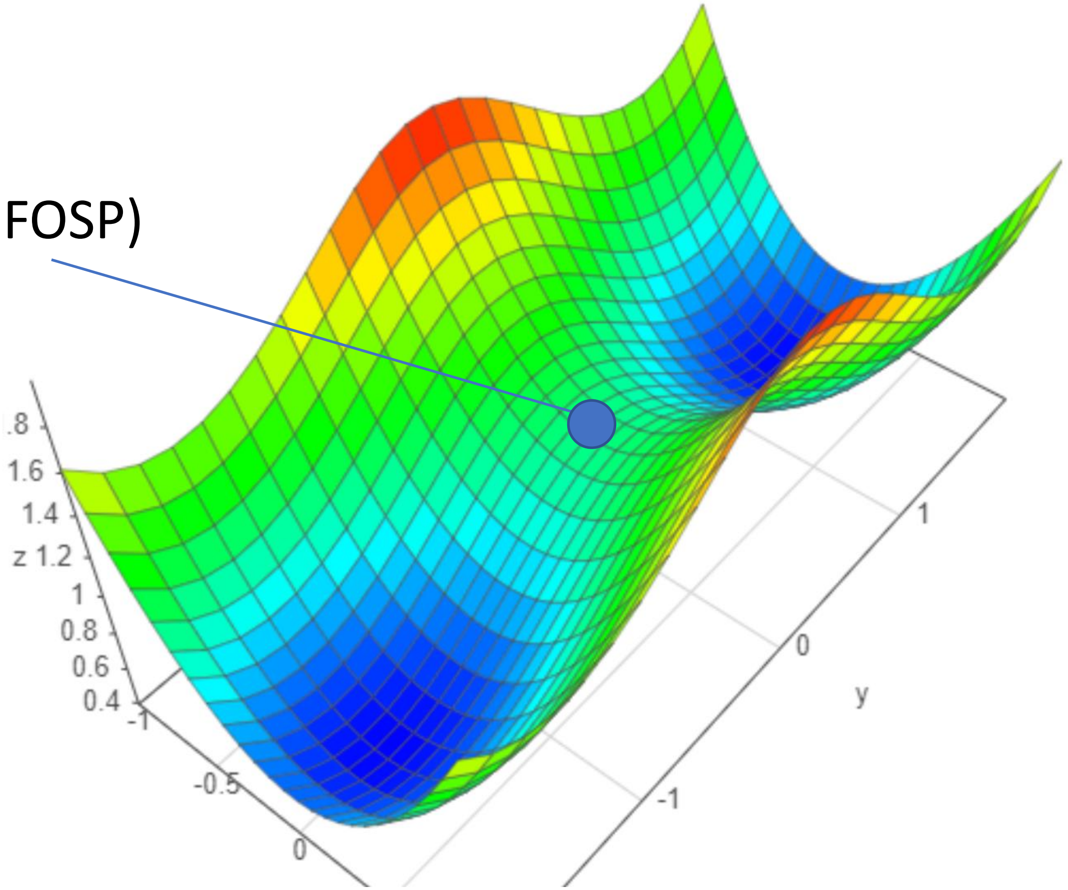


image credit: academo.org

Finding First Order Stationary Points

First Order Stationary Point (FOSP)



- Defined by: $\nabla f(w) = 0$

- But $\nabla^2 f(w)$ need not be positive semi-definite

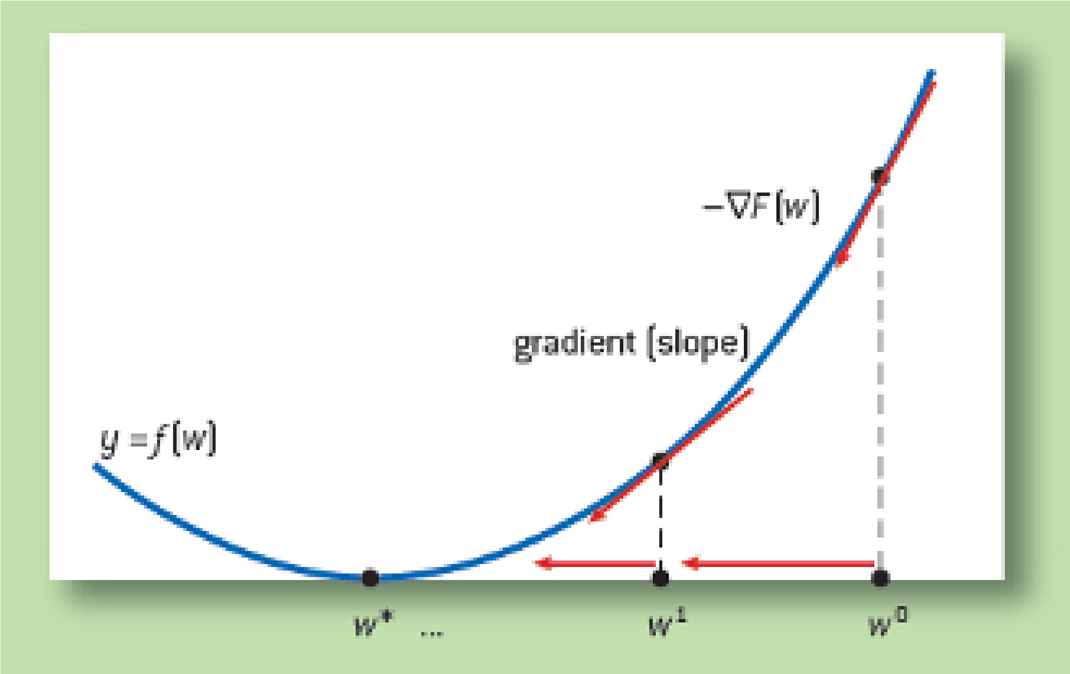
image credit: academo.org

Gradient Descent Methods

- Gradient descent:

- For t=1, 2, … (until convergence)
  - $w_{t+1} = w_t - \eta \nabla f(w_t)$

- η: step-size
- Assume:
  $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$



Convergence to FOSP

ⓐ $w_{t+1} = w_t - \eta \nabla f(w_t)$
ⓑ $\eta < \frac{1}{L}$

$$f(w_{t+1}) \le f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} ||w_{t+1} - w_t||^2$$

$$f(w_{t+1}) \le f(w_t) - \left(1 - \frac{L\eta}{2}\right) \eta ||\nabla f(w_t)||^2 \le f(w_t) - \frac{1}{2L} ||\nabla f(w_t)||^2$$

$$||\nabla f(w_t)||^2 \le f(w_t) - f(w_{t+1})$$

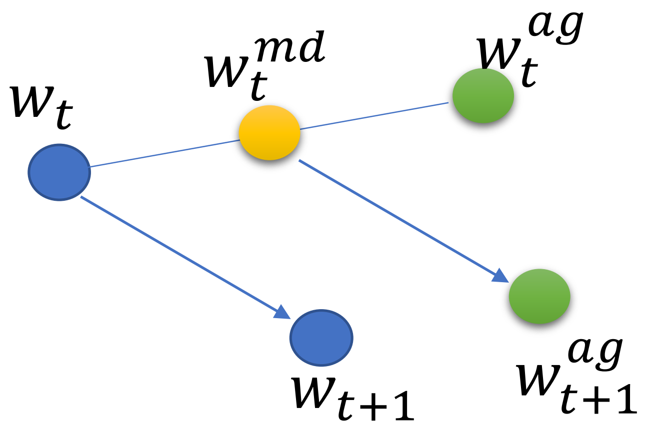
$$\frac{1}{2L} \sum_t ||\nabla f(w_t)||^2 \le f(w_0) - f(w_*)$$

$$\min_t ||\nabla f(w_t)|| \le \sqrt{\frac{2L\,(f(w_0) - f(w_*))}{T}} \le \epsilon$$

$$T = O\left(\frac{L \cdot (f(w_0) - f(w_*))}{\epsilon^2}\right)$$

Accelerated Gradient Descent for FOSP?

- For t=1, 2….T
  - $w_{t+1}^{md} = (1-\alpha_t)w_t^{ag} + \alpha_t w_t$
  - $w_{t+1} = w_t - \eta_t \nabla f(w_{t+1}^{md})$
  - $w_{t+1}^{ag} = w_t^{md} - \beta_t \nabla f(w_{t+1}^{md})$



- Convergence? $\min_t ||\nabla f(w_t)|| \leq \epsilon$
- For $T = O(\frac{\sqrt{L\cdot(f(w_0)-f(w_*))}}{\epsilon})$
- If convex: $T = O(\frac{(L\cdot(f(w_0)-f(w_*)))^{1/4}}{\sqrt{\epsilon}})$

Ghadimi and Lan - 2013

Non-convex Optimization: Sum of Functions

- What if the function has more structure?

$$\min_{w} \; f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

- $\nabla f(w) = \sum_{i=1}^{n} \nabla f_i(w)$
- I.e., computing gradient would require $O(n)$ computation

Does Stochastic Gradient Descent Work?

- For t=1, 2, … (until convergence)
  - Sample $i_t \sim Unif[1, n]$
  - $w_{t+1} = w_t - \eta \nabla f_{i_t}(w_t)$

Proof? $E_{i_t}[w_{t+1} - w_t] = \eta \nabla f(w_t)$

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} ||w_{t+1} - w_t||^2$$

$$\mathrm{E}[f(w_{t+1})] \leq E[f(w_t)] - \frac{\eta}{2} ||\nabla f(w_t)||^2 + \frac{L}{2} \eta^2 \cdot Var$$

$$\min_t ||\nabla f(w_t)|| \leq \frac{\left(L\left(f(w_0) - f(w_*)\right) \cdot Var\right)^{\frac{1}{4}}}{T^{\frac{1}{4}}} \leq \epsilon$$

$$T = O\left(\frac{L \cdot Var \cdot (f(w_0) - f(w_*))}{\epsilon^4}\right)$$

Summary: Convergence to FOSP

Algorithm	No. of Gradient Calls (Non-convex)	No. of Gradient Calls (Convex)
GD [Folkore; Nesterov]	$O\left(\frac{1}{\epsilon^2}\right)$	$O\left(\frac{1}{\epsilon}\right)$
AGD [Ghadimi & Lan-2013]	$O\left(\frac{1}{\epsilon}\right)$	$O\left(\frac{1}{\sqrt{\epsilon}}\right)$

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

Algorithm	No. of Gradient Calls	Convex Case
GD [Folkore]	$O(\frac{n}{\epsilon^2})$	$O(\frac{n}{\epsilon})$
AGD [Ghadimi & Lan’2013]	$O\left(\frac{n}{\epsilon}\right)$	$O\left(\frac{n}{\sqrt{\epsilon}}\right)$
SGD [Ghadimi & Lan’2013]	$O(\frac{1}{\epsilon^4})$	$O(\frac{1}{\epsilon^2})$
SVRG [Reddi et al-2016, Allen-Zhu&Hazan-2016]	$O(n + n^{\frac{2}{3}}/\epsilon^2)$	$O(n + \sqrt{n}/\epsilon^2)$
MSVRG [Reddi et al-2016]	$O(\min(\frac{1}{\epsilon^4}, \frac{n^{\frac{2}{3}}}{\epsilon^2}))$	$O\left(n + \frac{\sqrt{n}}{\epsilon^2}\right)$

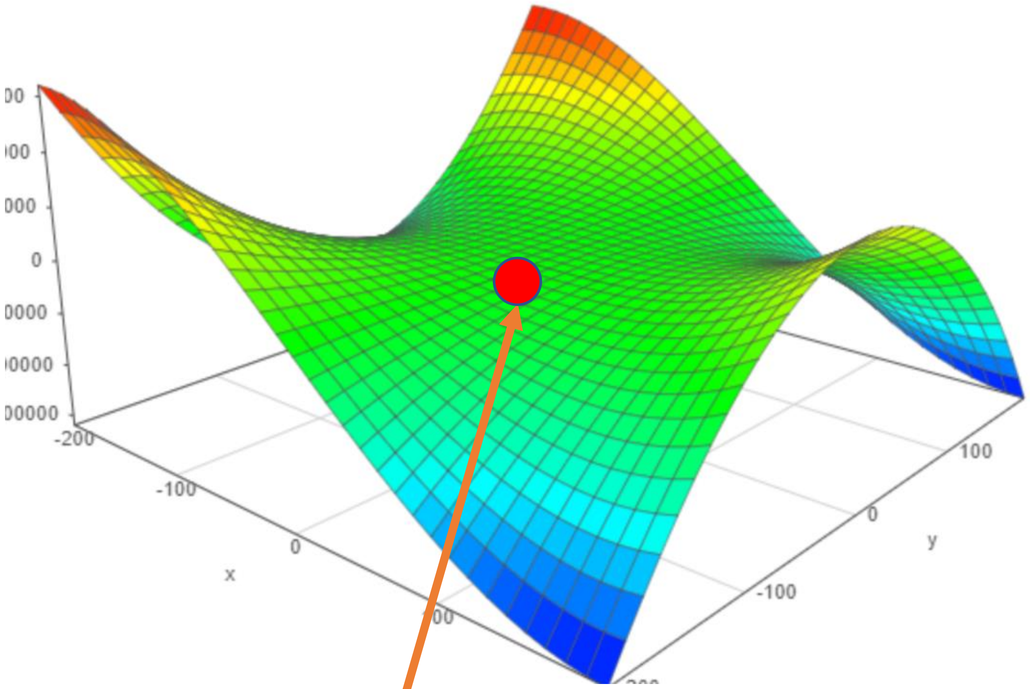
Finding Second Order Stationary Points (SOSP)

Second Order Stationary Point (SOSP) if:

- $\nabla f(w) = 0$
- $\nabla^2 f(w) \succcurlyeq 0$

Approximate SOSP:

- $||\nabla f(w)|| \leq \epsilon$
- $\lambda_{min}\left(\nabla^2 f(w)\right) \geq -\sqrt{\rho\epsilon}$



Second Order Stationary Point (SOSP)

image credit: academo.org

Cubic Regularization (Nesterov and Polyak-2006)

- For t=1, 2, … (until convergence)

$$w_{t+1} = \arg\min_{w} f(w_t) + \langle w - w_t, \nabla f(w_t)\rangle + \frac{1}{2}(w - w_t)^T \nabla^2 f(w_t)(w - w_t) + \frac{\rho}{6}||w - w_t||^3$$

- Assumption: Hessian continuity, i.e., $||\nabla^2 f(x) - \nabla^2 f(y)|| \leq \rho||x - y||$
- Convergence to SOSP? $T = O(\frac{1}{\epsilon^{1.5}})$
  - But requires Hessian computation! (even storage is $O(d^2)$)
  - Can we find SOSP using only gradients?

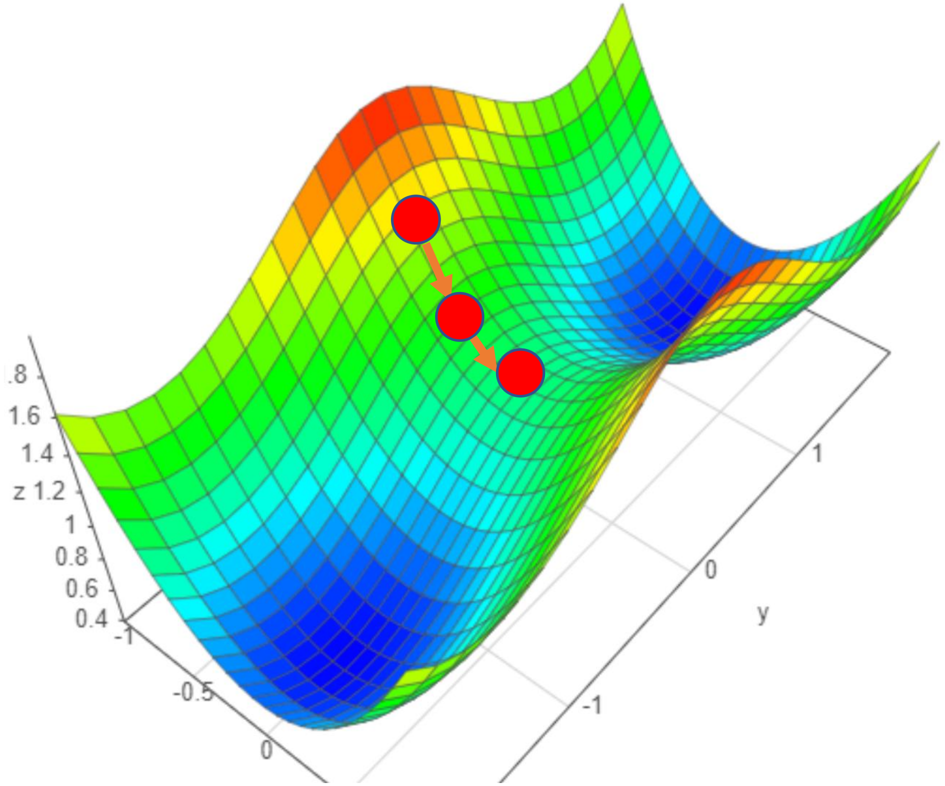
Noisy Gradient Descent for SOSP

- For t=1, 2, … (until convergence)
  - If ($||\nabla f(w_t|| \geq \epsilon$)
    - $w_{t+1} = w_t - \eta \nabla f(w_t)$
  - Else
    - $w_{t+1} = w_t + \zeta, \zeta \sim \gamma \cdot N(0, I)$
    - Update $w_{t+1} = w_t - \eta \nabla f(w_t)$ for next r iterations

- Claim: above algorithm converges to SOSP in $O(1/\epsilon^2)$

Ge et al-2015, Jin et al-2017

Proof



For t=1, 2, … (until convergence)

If ($||\nabla f(w_t|| \geq \epsilon$)

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$

Else

$$w_{t+1} = w_t + \zeta, \zeta \sim \gamma \cdot N(0, I)$$

Update $w_{t+1} = w_t - \eta \nabla f(w_t)$ for next r iterations

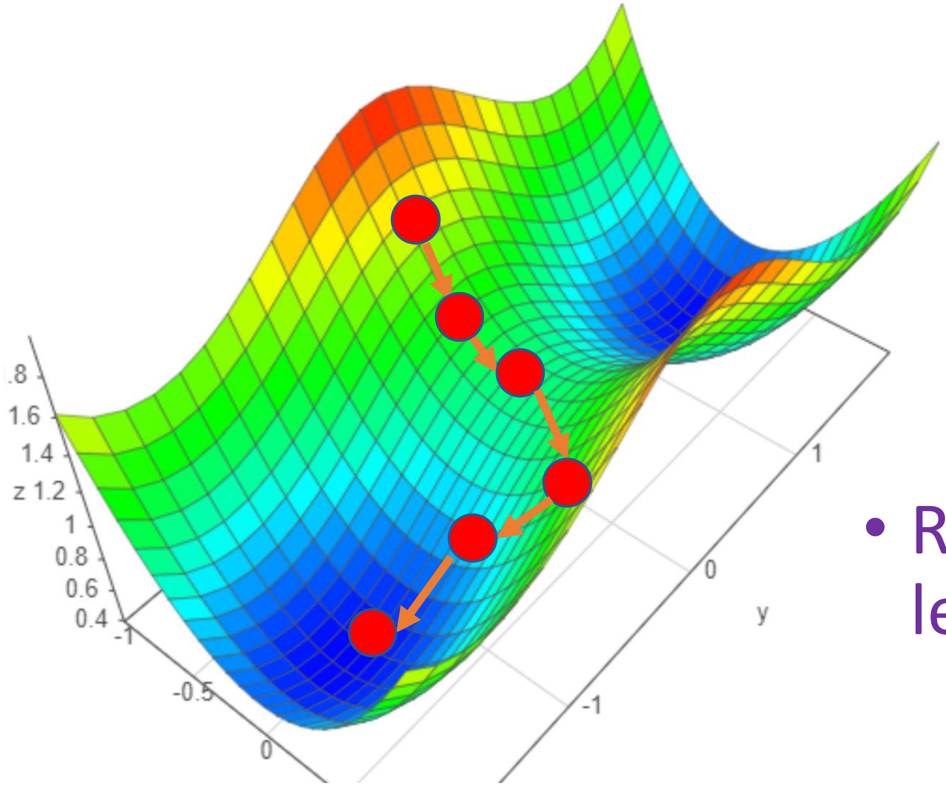
FOSP analysis: convergence in $O\left(\frac{1}{\epsilon^2}\right)$ iterations

But, $\nabla^2 f(w_t) \ngeq 0$

- That is, $\lambda_{min}\left(\nabla^2 f(w_t)\right) < -\sqrt{\rho\epsilon}$

image credit: academo.org

Proof



For t=1, 2, … (until convergence)

If ($||\nabla f(w_t|| \geq \epsilon$)

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$

Else

$$w_{t+1} = w_t + \zeta, \zeta \sim \gamma \cdot N(0, I)$$

Update $w_{t+1} = w_t - \eta \nabla f(w_t)$ for next r iterations

- Random perturbation with Gradient descent leads to decrease in objective function

image credit: academo.org

For t=1, 2, … (until convergence)

If ($||\nabla f(w_t|| \geq \epsilon$)

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$

Else

$$w_{t+1} = w_t + \zeta, \zeta \sim \gamma \cdot N(0, I)$$

Update $w_{t+1} = w_t - \eta \nabla f(w_t)$ for next r iterations

Proof?

- Random perturbation with Gradient descent leads to decrease in objective function
- Hessian continuity => function nearly quadratic in small neighborhood
- $f(w) \approx f(w_t) + \underbrace{\langle \nabla f(w_t), w - w_t \rangle}_{\to 0} + (w - w_t)^T \nabla^2 f(w_t)(w - w_t)$

$$w_{r+t} = w_{r-1+t} - \eta \nabla^2 f(w_t)(w_{r-1+t} - w_t)$$

$$\Rightarrow w_{r+t} - w_t = \underbrace{\left(I - \eta \nabla^2 f(w_t)\right)^r (w_{t+1} - w_t)}_{\text{Power Method}}$$

Proof?

For t=1, 2, … (until convergence)

If ($||\nabla f(w_t|| \geq \epsilon$)

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$

Else

$$w_{t+1} = w_t + \zeta, \zeta \sim \gamma \cdot N(0, I)$$

Update $w_{t+1} = w_t - \eta \nabla f(w_t)$ for next r iterations

- Random perturbation with Gradient descent leads to decrease in objective function
- Hessian continuity => function nearly quadratic in small neighborhood
- $f(w) \approx f(w_t) + \langle \nabla f(w_t), w - w_t \rangle + (w - w_t)^T \nabla^2 f(w_t)(w - w_t)$

$$w_{r+t} = w_{r-1+t} - \eta \nabla^2 f(w_t)(w_{r-1+t} - w_t)$$

$$\Rightarrow w_{r+t} - w_t = \left(I - \eta \nabla^2 f(w_t)\right)^r (w_{t+1} - w_t)$$

- $w_{r+t} - w_t$ converge to largest eigenvector of $I - \eta \nabla^2 f(w_t)$
  - Which is smallest (most negative) eigenvector of $\nabla^2 f(w_t)$
- Hence, $(w_{r+t} - w_t)^T \nabla^2 f(w_t)(w_{r+t} - w_t) \leq -\gamma^2 \sqrt{\rho\epsilon}$
- $f(w_{r+t}) \leq f(w_t) - \gamma^2 \sqrt{\rho\epsilon}$

Proof

For t=1, 2, … (until convergence)

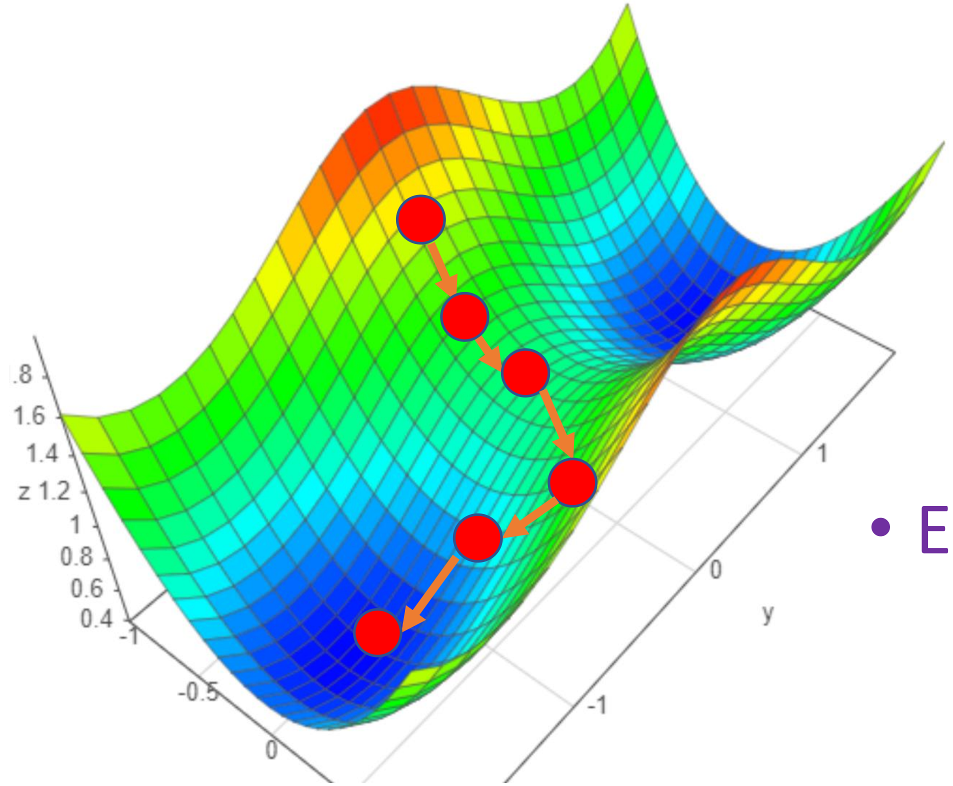
If ($||\nabla f(w_t|| \geq \epsilon$)

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$

Else

$$w_{t+1} = w_t + \zeta, \zeta \sim \gamma \cdot N(0, I)$$

Update $w_{t+1} = w_t - \eta \nabla f(w_t)$ for next r iterations



- Entrapment near SOSP

Final result: convergence to SOSP in $O(1/\epsilon^2)$

image credit: academo.org

Ge et al-2015, Jin et al-2017

Summary: Convergence to SOSP

Algorithm	No. of Gradient Calls (Non-convex)	No. of Gradient Calls (Convex)
Noisy GD [Jin et al-2017, Ge et al-2015]	$O\left(\frac{1}{\epsilon^2}\right)$	$O\left(\frac{1}{\epsilon}\right)$
Noisy Accelerated GD [Jin et al-2017]	$O\left(\frac{1}{\epsilon^{1.75}}\right)$	$O\left(\frac{1}{\sqrt{\epsilon}}\right)$
Cubic Regularization [Nesterov & Polyak-2006]	$O\left(\frac{1}{\epsilon^{1.5}}\right)$	N/A

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

Algorithm	No. of Gradient Calls	Convex Case
Noisy GD [Jin et al-2017, Ge et al-2015]	$O(\frac{n}{\epsilon^2})$	$O(\frac{n}{\epsilon})$
Noisy AGD [Jin et al-2017]	$O\left(\frac{n}{\epsilon^{1.75}}\right)$	$O\left(\frac{n}{\sqrt{\epsilon}}\right)$
Noisy SGD [Jin et al-2017, Ge et al-2015]	$O(\frac{1}{\epsilon^4})$	$O(\frac{1}{\epsilon^2})$
SVRG [Allen-Zhu-2018]	$O(n + n^{\frac{3}{4}}/\epsilon^2)$	$O(n + \sqrt{n}/\epsilon^2)$

Convergence to Global Optima?

- FOSP/SOSP methods can't even guarantee local convergence
- Can we guarantee global optimality for some "nicer" non-convex problems?
  - Yes!!!
  - Use statistics ☺

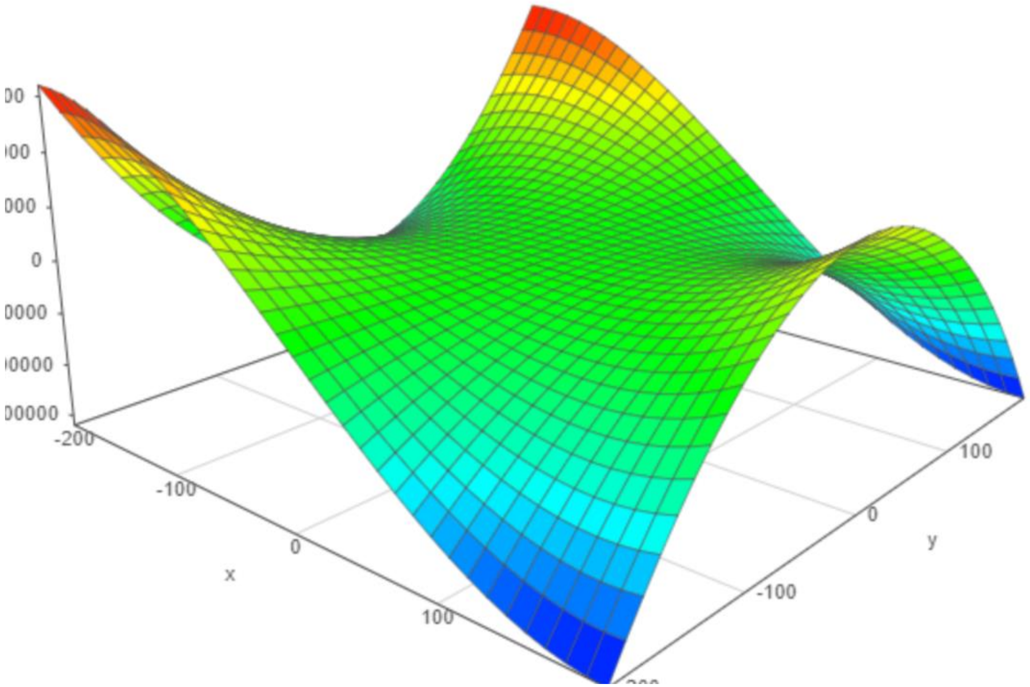


image credit: academo.org

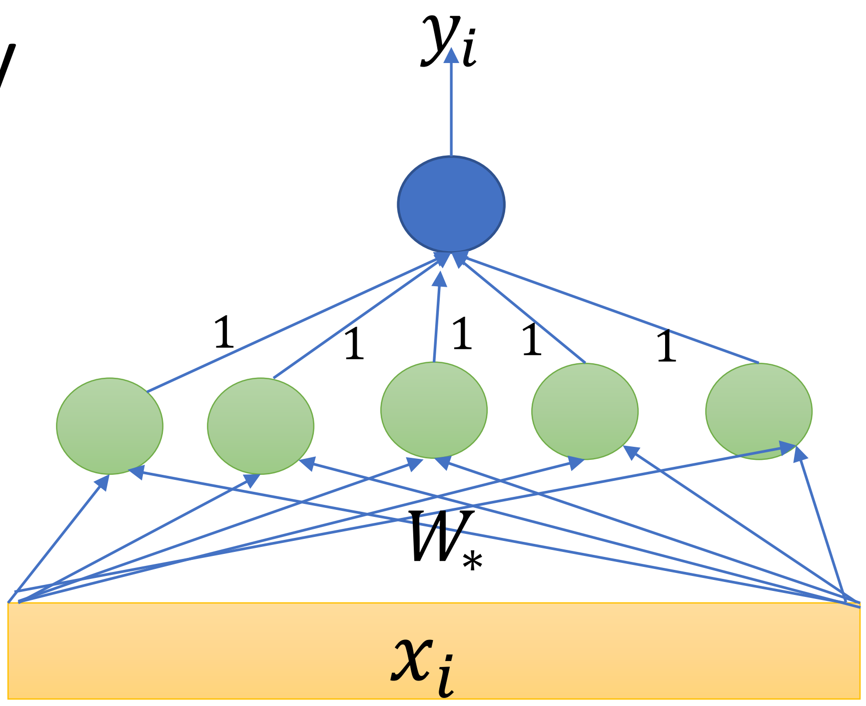
Can Statistics Help: Realizable models!

- Data points: $(x_i, y_i) \sim D$
- D: nice distribution
- $E[y_i] = \phi(x_i, w_*)$

$$\hat{w} = \arg\min_{w} \sum_{i} loss(y_i, \phi(x_i, w))$$

- That is, w_* is the optimal solution!
  - Parameter learning

Learning Neural Networks: Provably



- $y_i = 1 \cdot \sigma(W_* x_i)$
- $x_i \sim N(0, I)$

$$\min_W \sum_i \left(y_i - 1 \cdot \sigma(W x_i) \right)^2$$

- Does gradient descent converge to global optima: W_*?
  - NO!!!
  - The objective function has poor local minima [Shamir et al-2017, Lee et al-2017]

Learning Neural Networks: Provably

- But, no local minima within constant distance of W_*
- If,

$$||W_0 - W_*|| \leq c$$

Then, Gradient Descent ($W_{t+1} = W_t - \eta \nabla \mathrm{f}(\mathrm{W_t})$) converges to W_*

No. of iterations: $\log 1/\epsilon$

Can we get rid of initialization condition? Yes but by changing the network [Liang-Lee-Srikant'2018]

Zhong-Song-J-Bartlett-Dhillon'2017

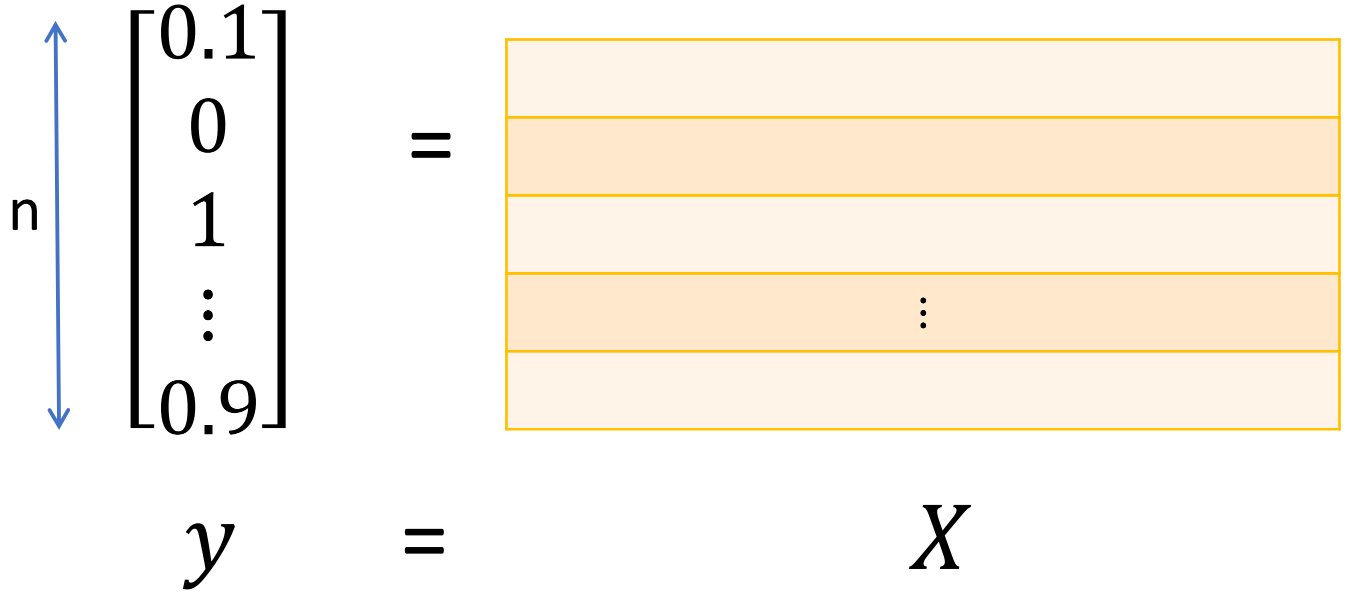
Learning with Structure

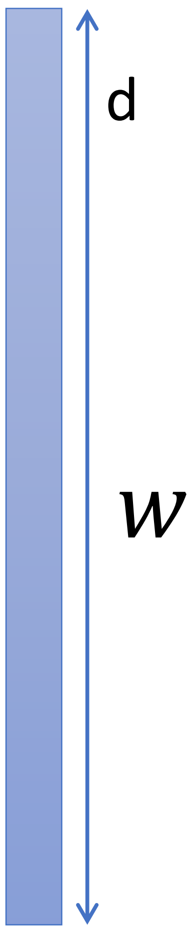
- $y_i = \phi(x_i, w_*),\ x_i \sim D \in R^d\ , \quad 1 \leq i \leq n$

- But no. of samples are limited!
  - For example, $if\ n \leq d$?

- Can we still recover w_*? In general, no!
  - But, what if w_* has some structure?

Sparse Linear Regression





- But: $n \ll d$
- w: s $-$sparse (s non-zeros)
  - Information theoretically: $n = s \log d$ samples should suffice

Learning with structure

$$\min_{w} f(w)$$
$$s.t.\ \ w \in C$$

- Linear classification/regression
  - $C = \{w,\ ||w||_0 \leq s\}$
  - $s \ll d$
- Matrix completion
  - $C = \{W, rank(W) \leq r\}$
  - $r \ll (d_1, d_2)$

Other Examples

- Low-rank Tensor completion
  - $C = \{W, \; tensor - rank(W) \leq r\}$
  - $r \ll (d_1, d_2, d_3)$

- Robust PCA
  - $C = \{W, W = L + S, rank(L) \leq r, ||S||_0 \leq s\}$
  - $r \ll (d_1, d_2), S \ll d_1 \times d_2$

Non-convex Structures

- Linear classification/regression
  - $C = \{w,\ ||w||_0 \leq s\}$
  - $s \ll d$

- NP-Hard
- $||w||_0$: Non-convex

- Matrix completion
  - $C = \{W, rank(W) \leq r\}$
  - $r \ll (d_1, d_2)$

- NP-Hard
- $rank(W)$: Non-convex

Non-convex Structures

- Low-rank Tensor completion
  - $C = \{W,\ tensor - rank(W) \leq r\}$
  - $r \ll (d_1, d_2, d_3)$

- Indeterminate
- $tensorrank(W)$: Non-convex

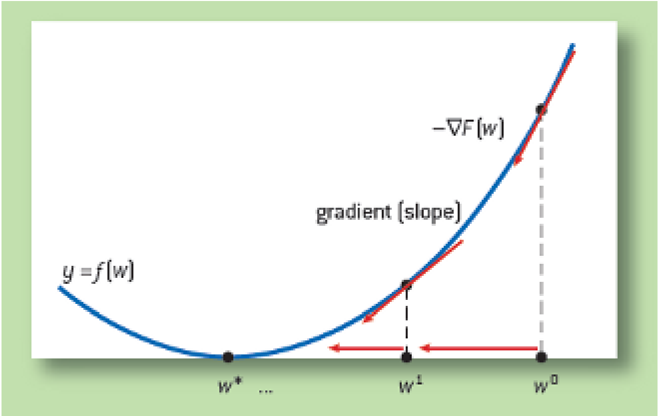
- Robust PCA
  - $C = \{W, W = L + S, rank(L) \leq r, ||S||_0 \leq s\}$
  - $r \ll (d_1, d_2), S \ll d_1 \times d_2$

- NP-Hard
- $rank(W), ||S||_0$: Non-convex

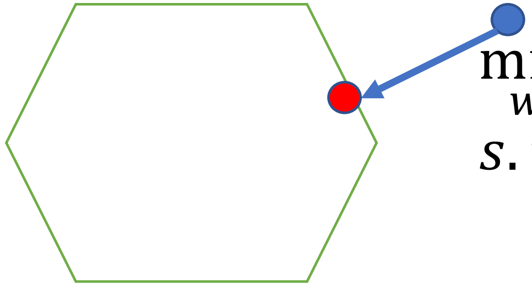
Technique: Projected Gradient Descent

$$\min_{w} f(w)$$
$$s.t.\ \ w \in C$$

- $w_{t+1} = w_t - \nabla_w f(w_t)$



- $w_{t+1} = P_C(w_{t+1})$



$$\min_{w} ||w - w_{t+1}||^2$$
$$s.t.\ \ w \in C$$

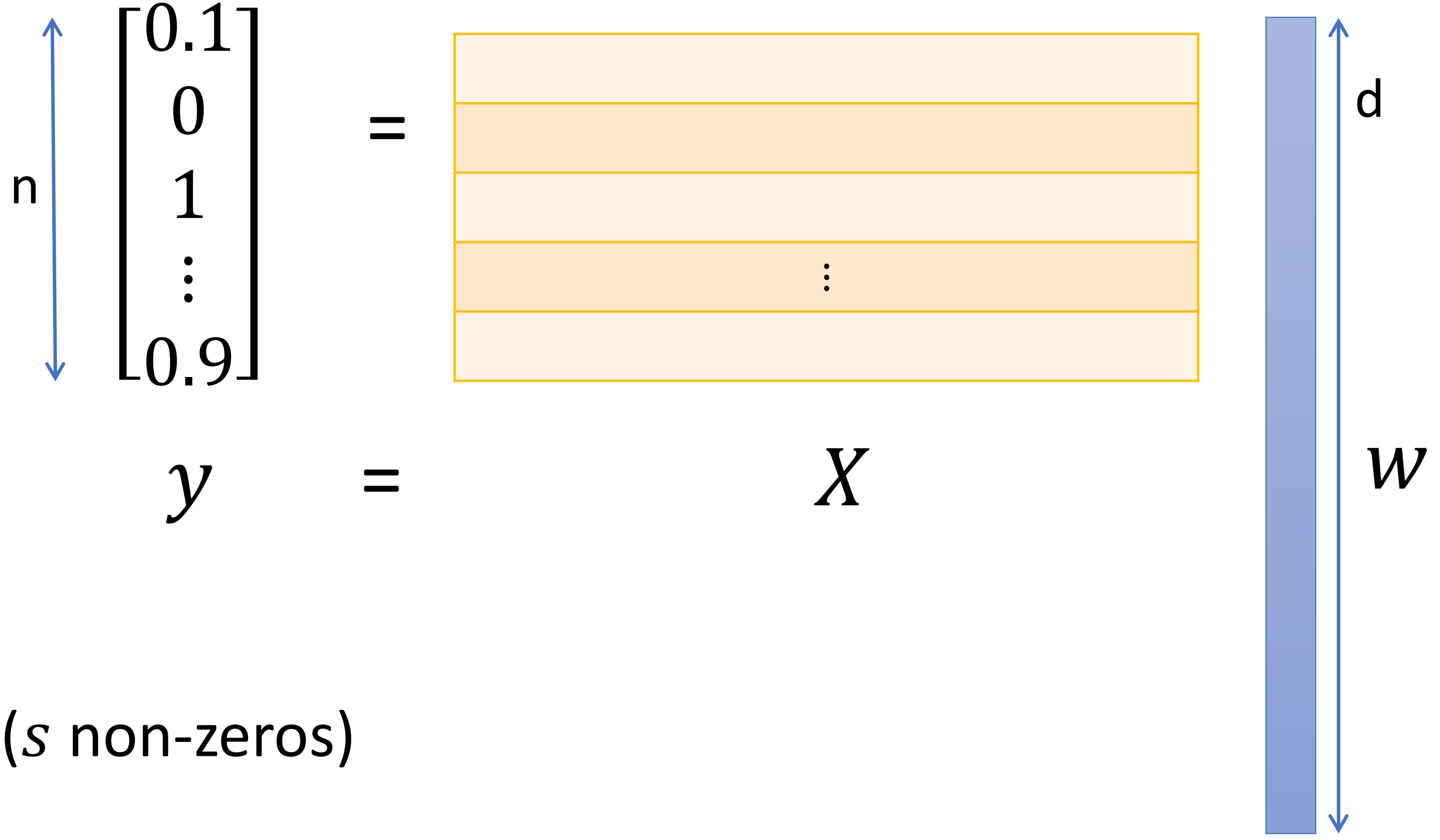
Results for Several Problems

- Sparse regression [Jain et al.'14, Garg and Khandekar'09]
  - Sparsity
- Robust Regression [Bhatia et al.'15]
  - Sparsity+output sparsity
- Vector-value Regression [Jain & Tewari'15]
  - Sparsity+positive definite matrix
- Dictionary Learning [Agarwal et al.'14]
  - Matrix Factorization + Sparsity
- Phase Sensing [Netrapalli et al.'13]
  - System of Quadratic Equations

Results Contd…

- Low-rank Matrix Regression [Jain et al.’10, Jain et al.’13]
  - Low-rank structure
- Low-rank Matrix Completion [Jain & Netrapalli’15, Jain et al.’13]
  - Low-rank structure
- Robust PCA [Netrapalli et al.’14]
  - Low-rank ∩ Sparse Matrices
- Tensor Completion [Jain and Oh’14]
  - Low-tensor rank
- Low-rank matrix approximation [Bhojanapalli et al.’15]
  - Low-rank structure

Sparse Linear Regression



- But: $n \ll d$
- w: s −sparse (s non-zeros)

Sparse Linear Regression

$$\min_{w} ||y - Xw||^2$$
$$s.t. \;\; ||w||_0 \leq s$$

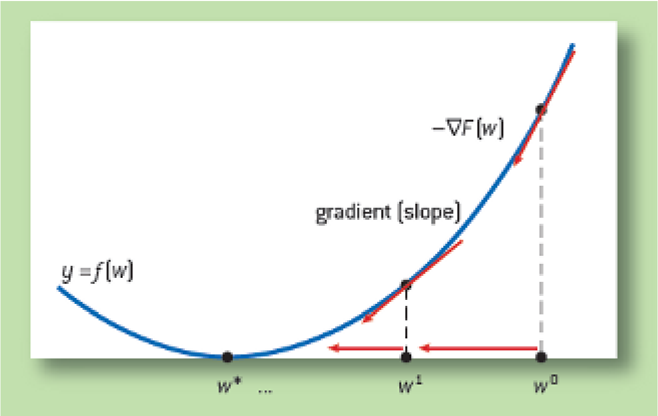
- $||y - Xw||^2 = \sum_i (y_i - \langle x_i, w \rangle)^2$
- $||w||_0$: number of non-zeros

- NP-hard problem in general ☹
  - L_0: non-convex function

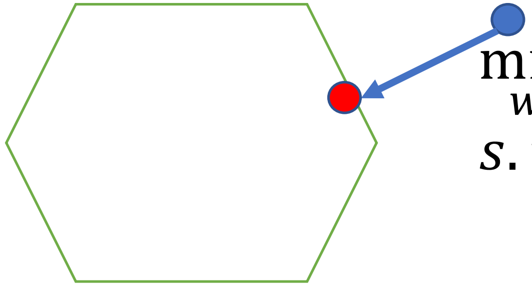
Technique: Projected Gradient Descent

$$\min_{w} f(w) = ||y - Xw||^2$$
$$s.t.\ \ ||w||_0 \leq s$$

- $w_{t+1} = w_t - \nabla_w f(w_t)$



- $w_{t+1} = P_s(w_{t+1})$



$$\min_{w} ||w - w_{t+1}||^2$$
$$s.t.\ \ ||w||_0 \leq s$$

[Jain, Tewari, Kar'2014]

Statistical Guarantees

$$y_i = \langle x_i, w^* \rangle + \eta_i$$

- $x_i \sim N(0, \Sigma)$
- $\eta_i \sim N(0, \zeta^2)$
- $w^*: s$ −sparse

$$|| \widehat{w} - w^* || \leq \frac{\zeta \kappa^3 \sqrt{s \log d}}{\sqrt{n}}$$

- $\kappa = \lambda_1(\Sigma) / \lambda_d(\Sigma)$

[Jain, Tewari, Kar’2014]

Low-rank Matrix Completion

users

movies	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3			5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

- unknown rating   - rating between 1 to 5

$$\min_{W} \sum_{(i,j)\in\Omega} \left(W_{ij} - M_{ij}\right)^2$$

$$s.t \qquad \mathbf{rank}(W) \le r$$

Ω: set of known entries

- Special case of low-rank matrix regression
- However, assumptions required by the regression analysis not satisfied

Technique: Projected Gradient Descent

- $W_0 = 0$
- For t=0:T-1

$$W_{t+1} = P_r(W_t - \eta \nabla f(W_t))$$

- $P_k(Z)$: projection onto set of rank-r projection
- Singular Value Projection
- Pros:
  - Fast (always, rank-r SVD)
  - Matrix completion: $O(d \cdot r^3)$!
- Cons: In general, might not even converge
- Our Result: Convergence under "certain" assumptions

[Jain, Tewari, Kar'2014], [Netrapalli, Jain'2014], [Jain, Meka, Dhillon'2009]

Guarantees

- Projected Gradient Descent:
  - $W_{t+1} = P_r\big(W_t - \eta \nabla_W f(W_t)\big), \qquad \forall t$
- Show ϵ-approximate recovery in $\log \frac{1}{\epsilon}$ iterations
- Assuming:
  - M: incoherent
  - Ω: uniformly sampled
  - $|\Omega| \geq n \cdot r^5 \cdot \log^3 n$
- First near linear time algorithm for **exact** Matrix Completion with finite samples

[J., Netrapalli'2015]

General Result for Any Function

$$\min_{w} f(w)$$
$$s.t.\ \ w \in C$$

- $f: R^d \to R$
- f: satisfies RSC/RSS, i.e.,

$$\alpha \cdot I_{d\times d} \preccurlyeq H(w) \preccurlyeq L \cdot I_{d\times d}, \qquad if, w \in C$$

- PGD guarantee: $$f(w_T) \leq f(w^*) + \epsilon$$

After $T = O(\log\left(\frac{f(w^0)}{\epsilon}\right))$ steps

- If $\frac{L}{\alpha} \leq 1.5$

[J., Tewari, Kar'2014]

Learning with Latent Variables

$$\min_{w,z} f(w, z)$$

- Typically, z are latent variables
- E.g., clustering: w: means of clusters, z: cluster index

- $f: \mathrm{non} - \mathrm{convex}$
  - NP-hard to solve in general

Alternating Minimization

$$z_{t+1} = \arg\min_{z} f(w_t, z)$$
$$w_{t+1} = \arg\min_{w} f(w, z_{t+1})$$

- For example, if $f(w_t, z)$ is convex and $f(w, z_t)$ is convex
- Does that imply $f(w, z)$ is convex?
  - No!!!
  - $f(w, z) = w \cdot z$
  - Linear in both w, z individually
- So can Alt. Min. converge to global optima?

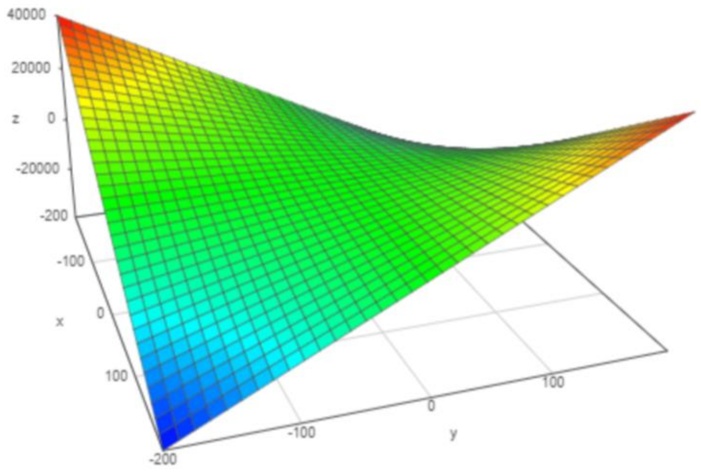


image credit: academo.org

Low-rank Matrix Completion

users

movies	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3			5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

- unknown rating - rating between 1 to 5

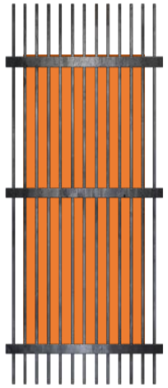
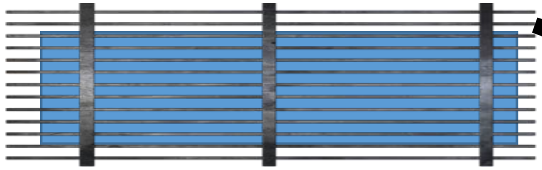
$$\min_{W} \sum_{(i,j)\in\Omega} \left(W_{ij} - M_{ij}\right)^2$$

$$s.t \qquad \mathbf{rank}(W) \leq r$$

Ω: set of known entries

- Special case of low-rank matrix regression
- However, assumptions required by the regression analysis not satisfied

Matrix Completion: Alternating Minimization

$\Big\|\ \mathbf{y} - \mathbf{X} \cdot \Big($  $\times$  $\Big)\Big\|_{\mathbf{F}}^{\mathbf{2}}$

$$W \cong U \times V^T$$

$$V^{t+1} = \min_V ||y - X \cdot (U^t V^T)\,||_2^2$$

$$U^{t+1} = \min_U ||y - X \cdot \left(U(V^{t+1})^T\right)||_2^2$$

Results: Alternating Minimization

- Provable global convergence [J., Netrapalli, Sanghavi'13]
- Rate of convergence: geometric

$$||W_T - W^*|| \leq 2^{-T}$$

- Assumptions:
  - Matrix regression: RIP
  - Matrix completion: uniform sampling and no. samples $|\Omega| \geq O(dk^6)$

[Jain, Netrapalli, Sanghavi'13]

General Results

$$\min_{w,z} f(w,z)$$

- Alternating minimization: optimal?
- If:
  - Joint Restricted Strong Convexity (Strong convexity close to the optimal)
  - Restricted Smoothness (smoothness near optimal)
  - Cross-product bound:

$$|\langle w - w_*, \nabla_w f(w,z) - \nabla_w f(w,z_*)\rangle - \langle z - z_*, \nabla_z f(w,z) - \nabla_z f(w_*,z)\rangle| \leq O(|w - w_*|^2 + |z - z_*|^2)$$

Ha and Barber-2017, Jain and Kar-2018

Summary I

Non-convex Optimization: two approaches

1. General non-convex functions
   a. First Order Stationary Point
   b. Second Order Stationary Point
2. Statistical non-convex functions: learning with structure
   a. Projected Gradient Descent (RSC/RSS)
   b. Alternating minimization/EM algorithms (RSC/RSS)

Summary II

- First Order Stationary Point :$f(w) \leq f(w') + ||w - w'||^2$
  - Tools: gradient descent, acceleration, stochastic gd, variance reduction
  - Key quantity: iteration complexity
  - Several questions: for example, can we do better? Especially in finite sum setting

- Second order stationary point: $f(w) \leq f(w') + ||w - w'||^3$
  - Tools: noise+gd, noise+acceleration, noise+sgd, noise+variance reduction
  - Several questions: better rates? Can we remove Lipschitz condition on Hessian?

Summary III

- Projected Gradient Descent
  - Works under statistical conditions like RSC/RSS
  - Still several open questions for most problems
  - E.g., tight guarantees support recovery for sparse linear regression?

- Alternating minimization
  - Works under some assumptions on f
  - What is the weakest condition on f for Alt. Min. to work?