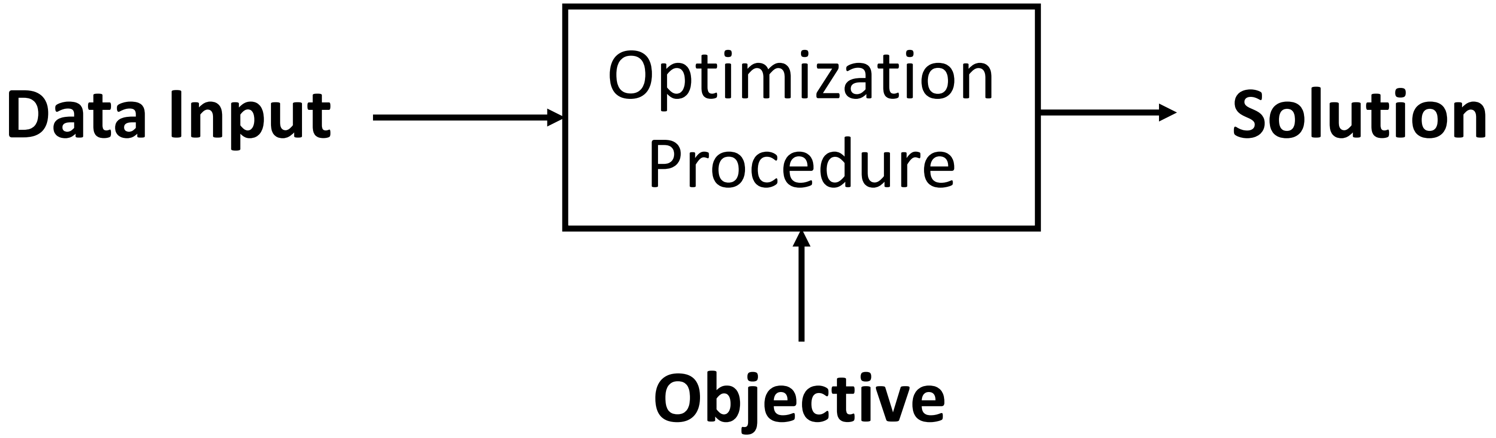
An Analytical Formula of Population Gradient for two-layered ReLU network and its Applications in Convergence and Critical Point Analysis

Yuandong Tian

Facebook AI Research

ICML 2017

General Framework of Optimization



Optimization with Guarantees

Data Input
Almost Any

→ Optimization Procedure → **Solution**

Many guarantees

Objective
Convex Objective
Matrix Factorization
Eigenvalue decomposition
Linear programming
…

Optimization we want to understand

Data Input
Almost Any

→ Optimization Procedure → **Solution**

Objective
Non-Convex Objective
(e.g, neural networks)

No guarantees
NP-hard

How about modeling data distribution?

Data Input → Optimization Procedure → **Solution**

~~Almost Any~~

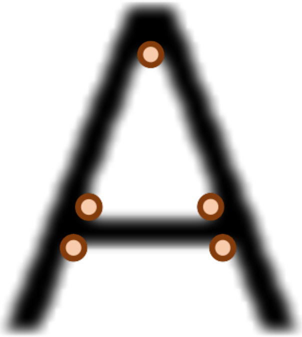
Specific distribution

Objective

Non-Convex Objective
(e.g, neural networks)

	Simple Objective	Complicated Objective
General Data Distribution	Many Guarantees	Not feasible/NP-hard
Specific Data Distribution		?

Example: Image Alignment



Template I_0

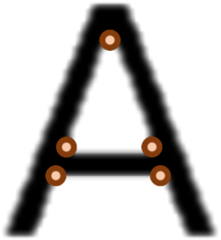
Deformed image $I_{\mathbf{p}^*}$

Landmark estimation $\mathbf{p}$

A Globally Optimal Data-Driven Approach for Image Distortion Estimation, Y. Tian and S. Narasimhan, CVPR 2010

Hierarchical Data-Driven Descent for Efficient Optimal Deformation Estimation, Y. Tian and S. Narasimhan, ICCV 2013

Example: Image Alignment



Template I_0

Deformed image $I_{\mathbf{p}^*}$

Landmark estimation $\mathbf{p}$

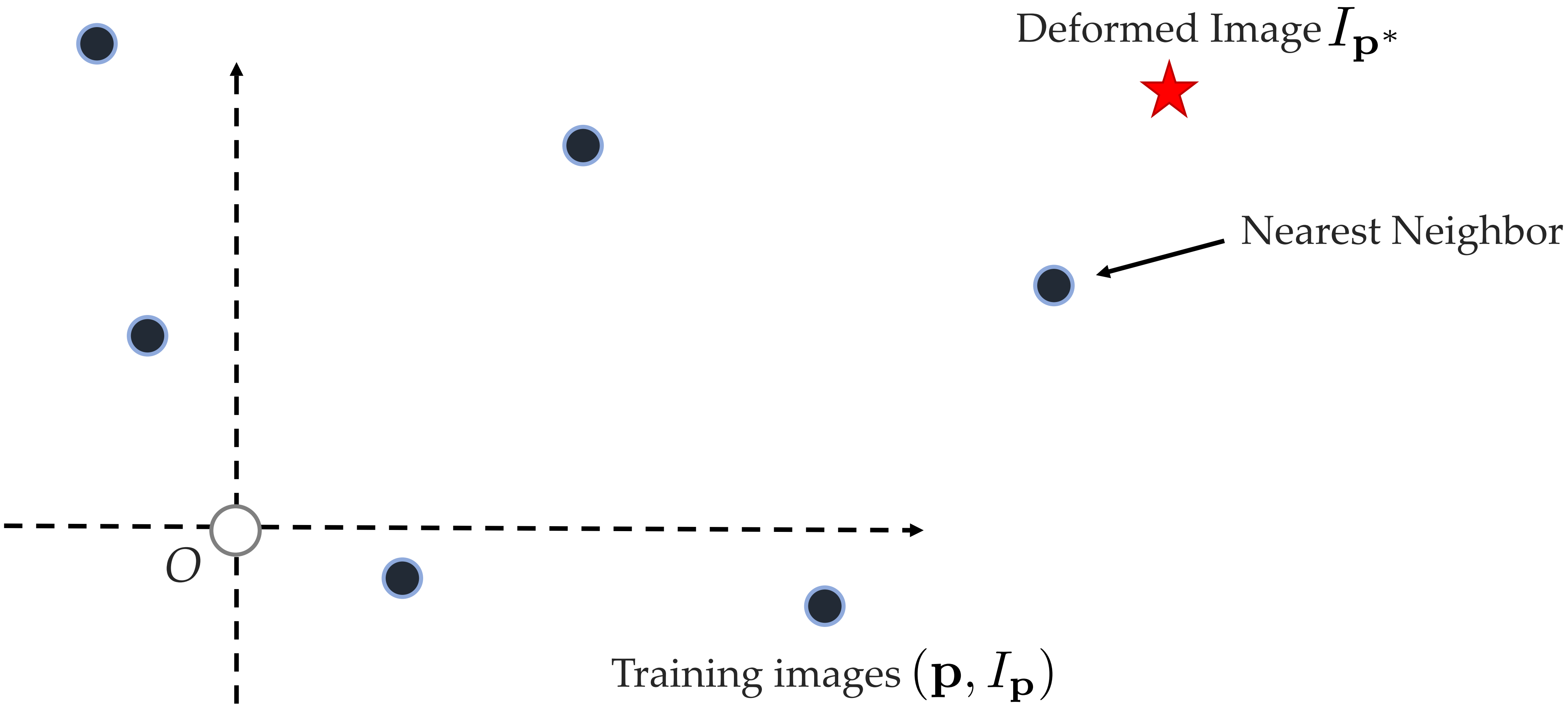
To achieve $\|\mathbf{p} - \mathbf{p}^*\| \leq \epsilon$:

Method	Sample complexity
Gradient descent	Local optimality
Nearest Neighbor	$O(1/\epsilon^d)$

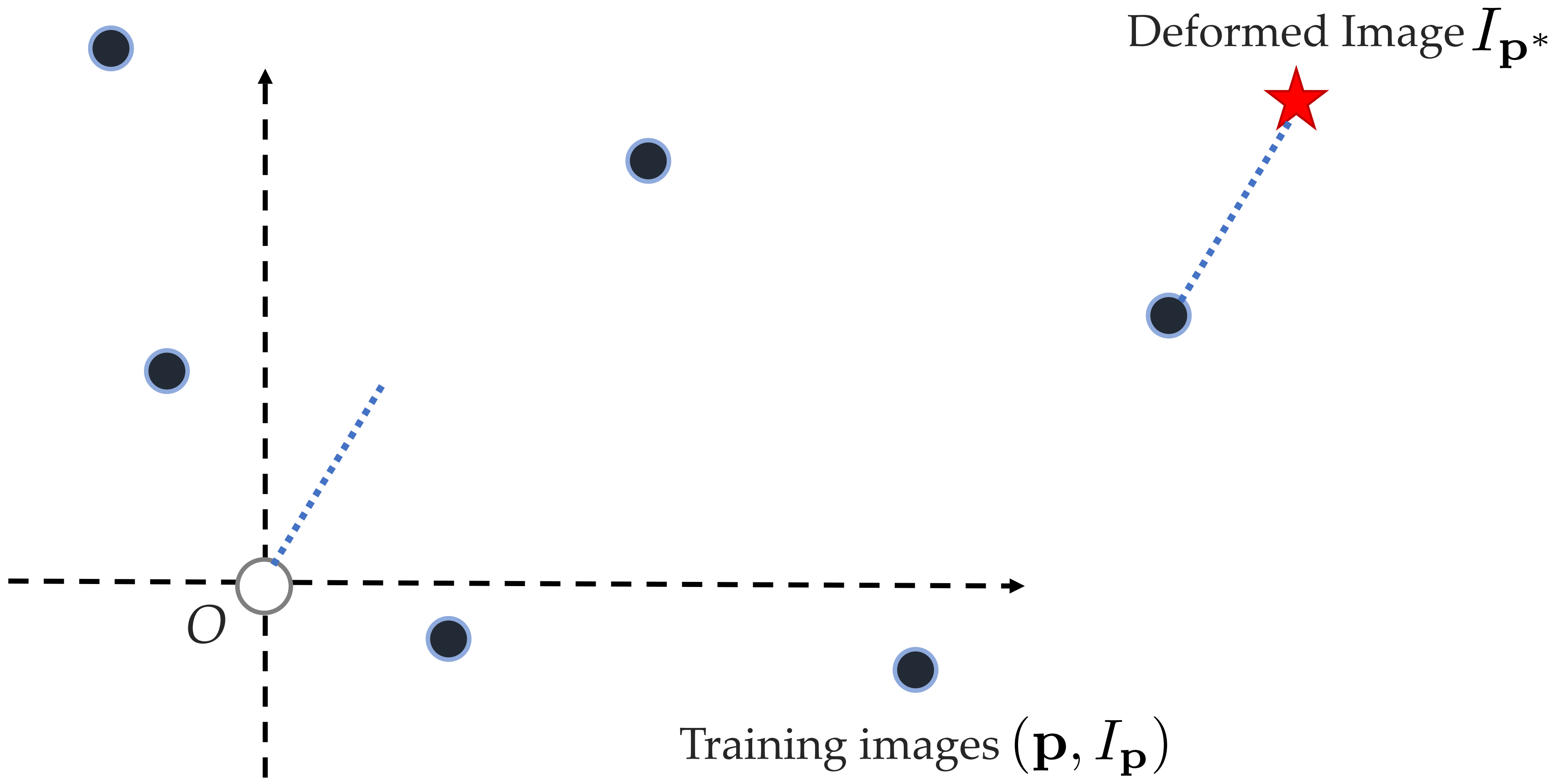
<u>Non-convex objective</u>

$$\min_{\mathbf{p}} J(\mathbf{p}; I_{\mathbf{p}^*}) = \min_{\mathbf{p}} \|I_{\mathbf{p}^*}(W(\cdot;\mathbf{p}) - I_0(\cdot)\|^2$$

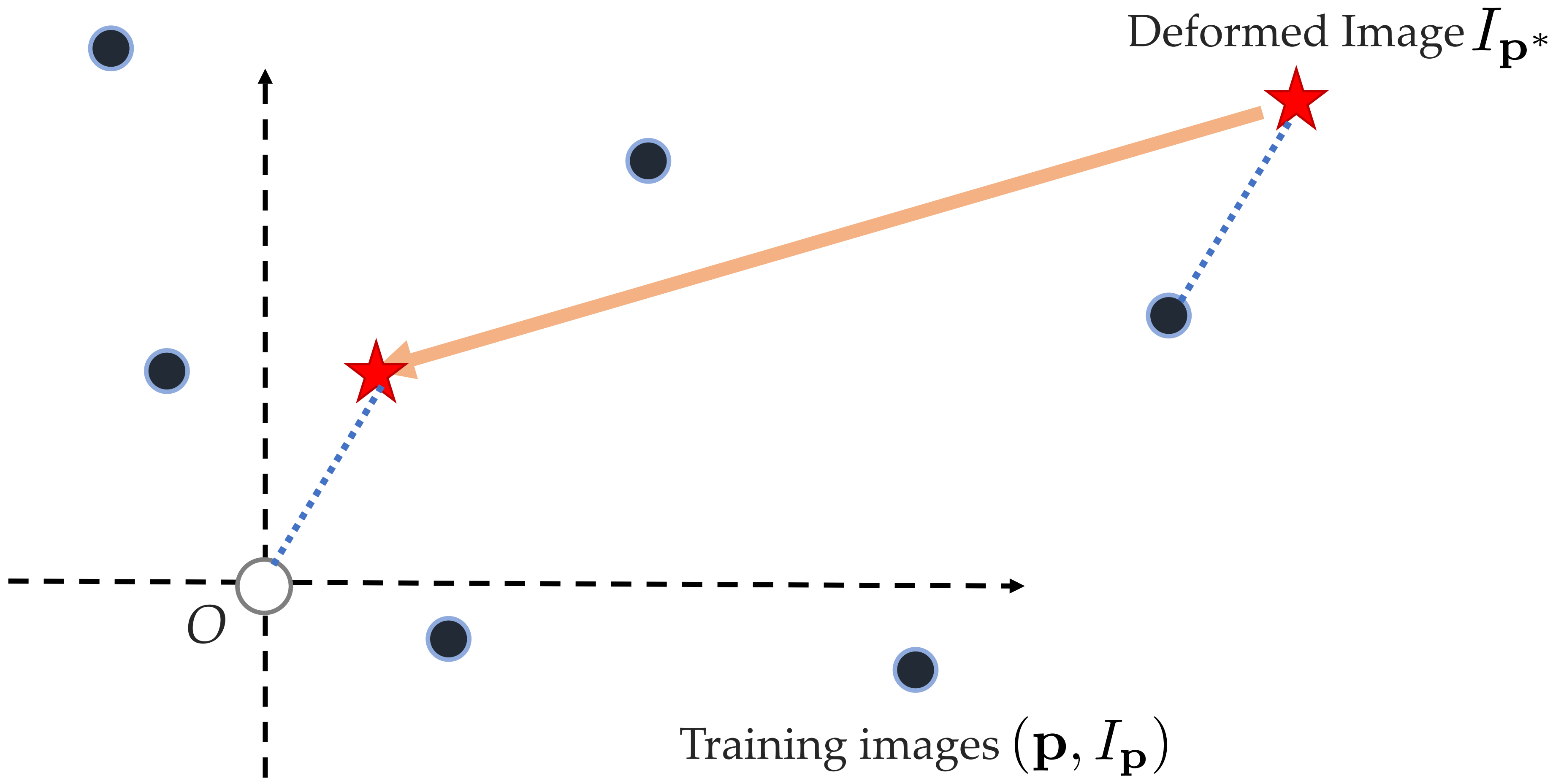
Example: Data-Driven Descent



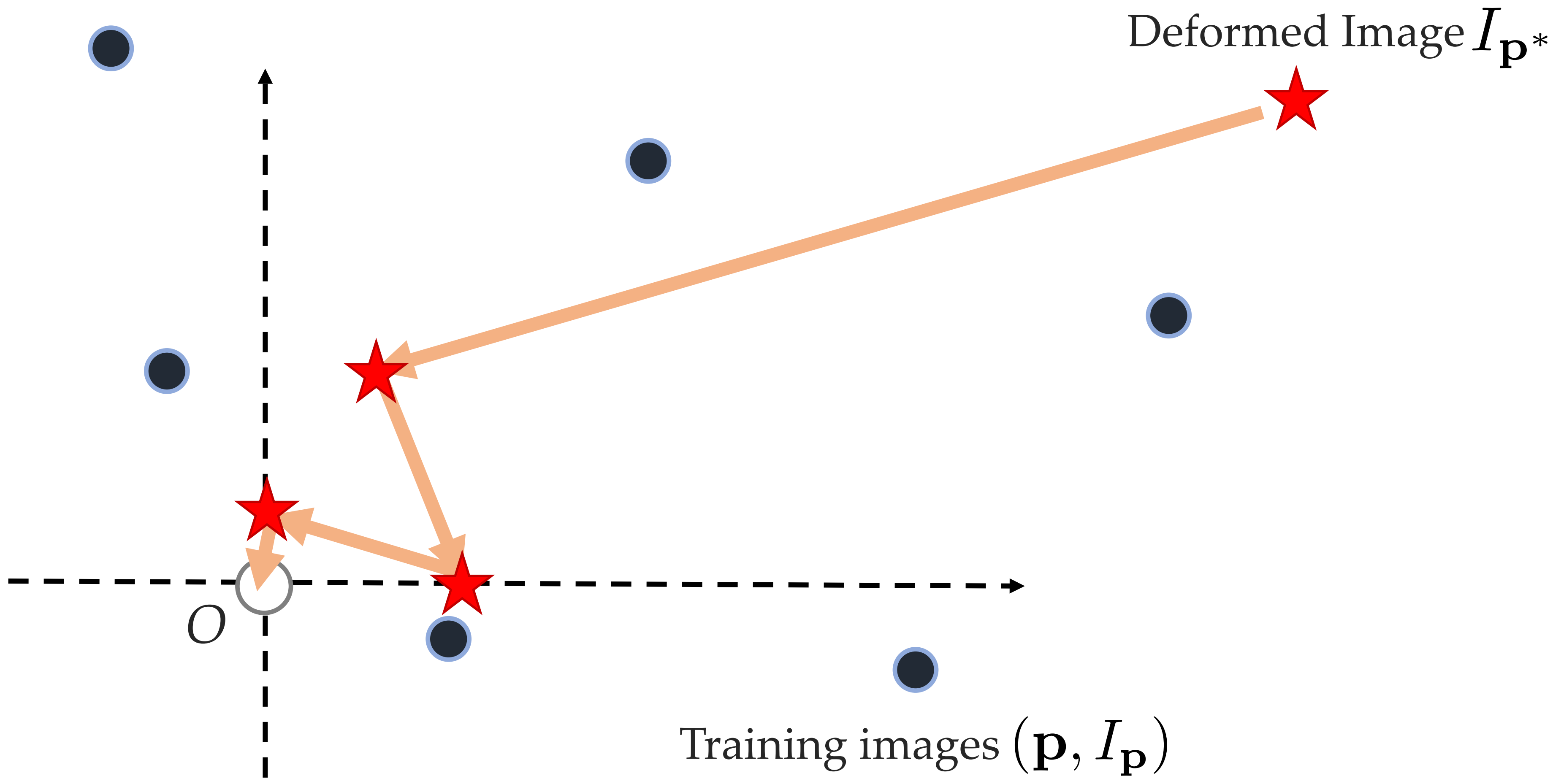
Example: Data-Driven Descent



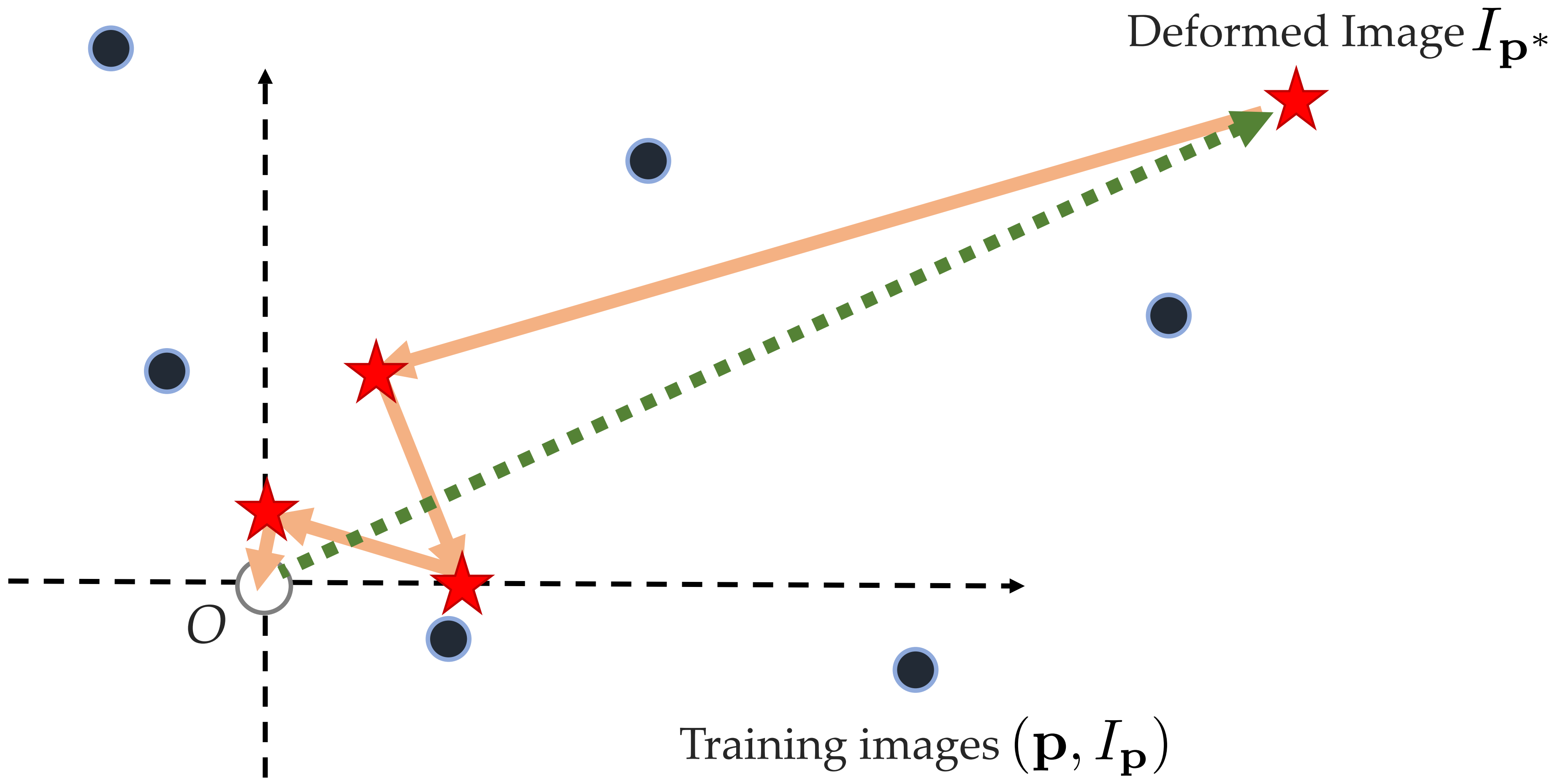
Example: Data-Driven Descent



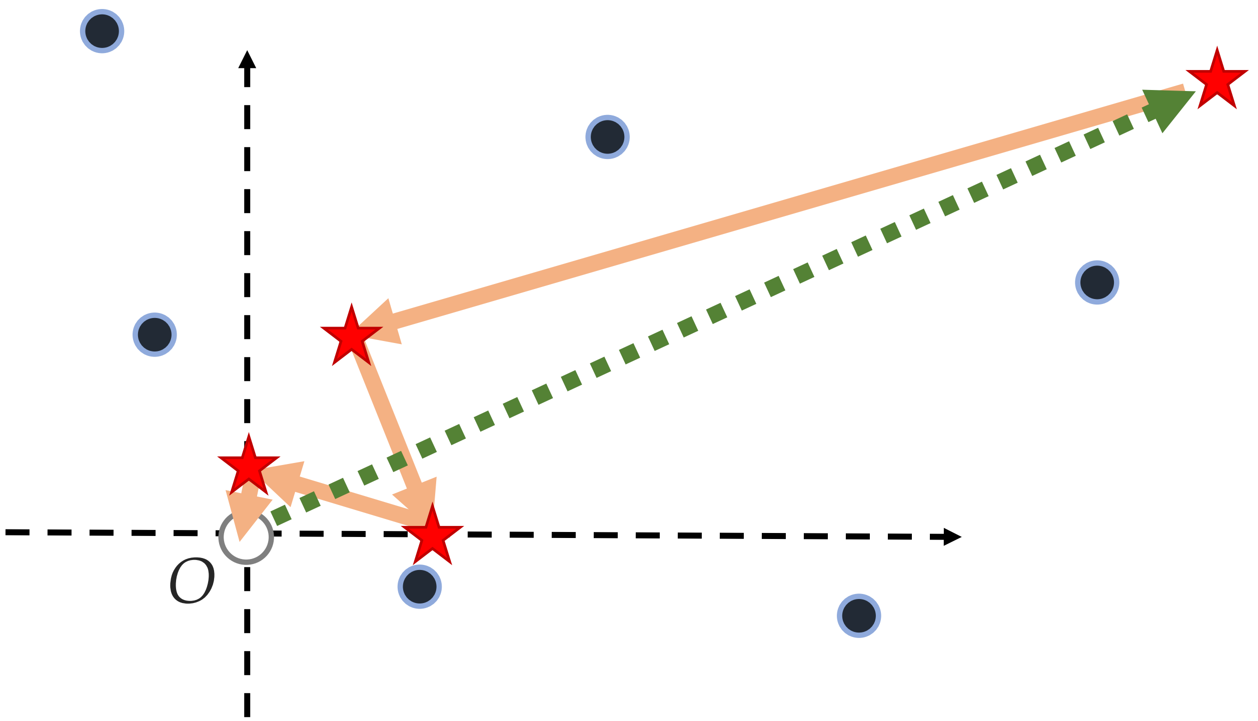
Example: Data-Driven Descent



Example: Data-Driven Descent



Example: Data-Driven Descent

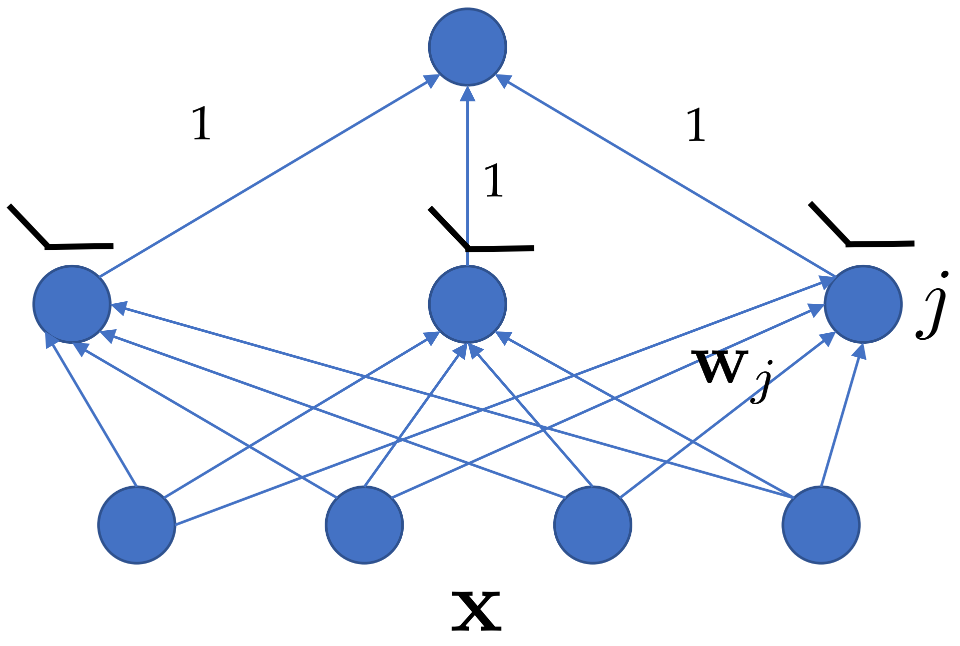


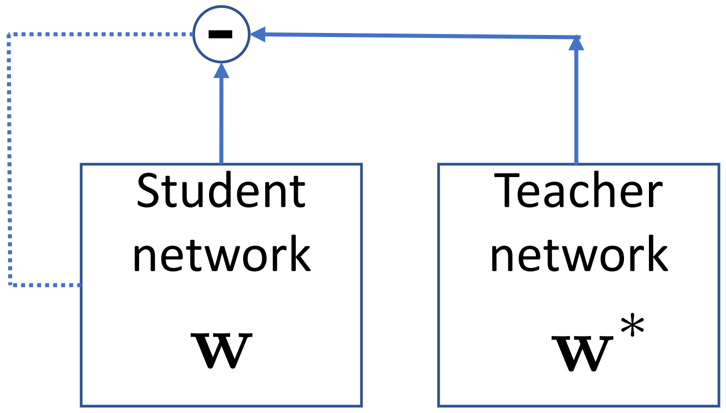
To achieve $\|\mathbf{p} - \mathbf{p}^*\| \leq \epsilon$:

Method	Sample complexity
Gradient descent	Local optimality
Nearest Neighbor	$O(1/\epsilon^d)$
[Y. Tian and S. Narasimhan, CVPR 10]	$O(C^d \log 1/\epsilon)$
[Y. Tian and S. Narasimhan, ICCV 13]	$O(C_1^d + C_2 \log 1/\epsilon)$

How about deep models?

This paper





$$g(\mathbf{x}; \mathbf{w}) = \sum_{j=1}^{K} \sigma(\mathbf{w}_j^T \mathbf{x}) \qquad \sigma(x) = \max(x, 0)$$

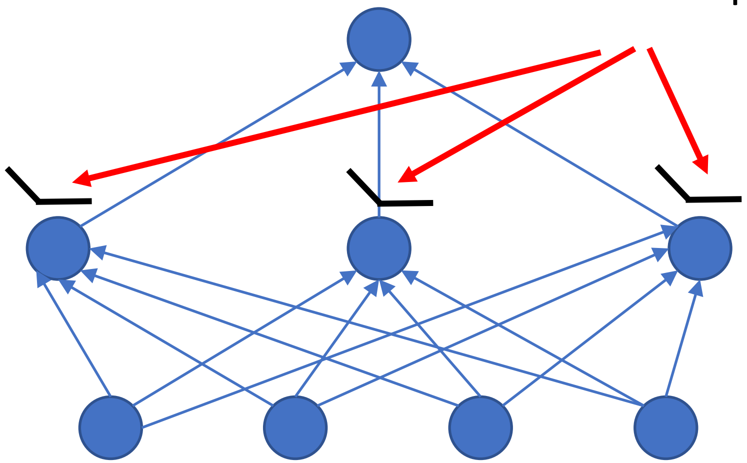
$$\mathbb{E}_{\mathbf{x}} J(\mathbf{x}; \mathbf{w}) = \frac{1}{2} \mathbb{E}_{\mathbf{x}} \left[\| g(\mathbf{x}; \mathbf{w}^*) - g(\mathbf{x}; \mathbf{w}) \|^2 \right]$$

population objective

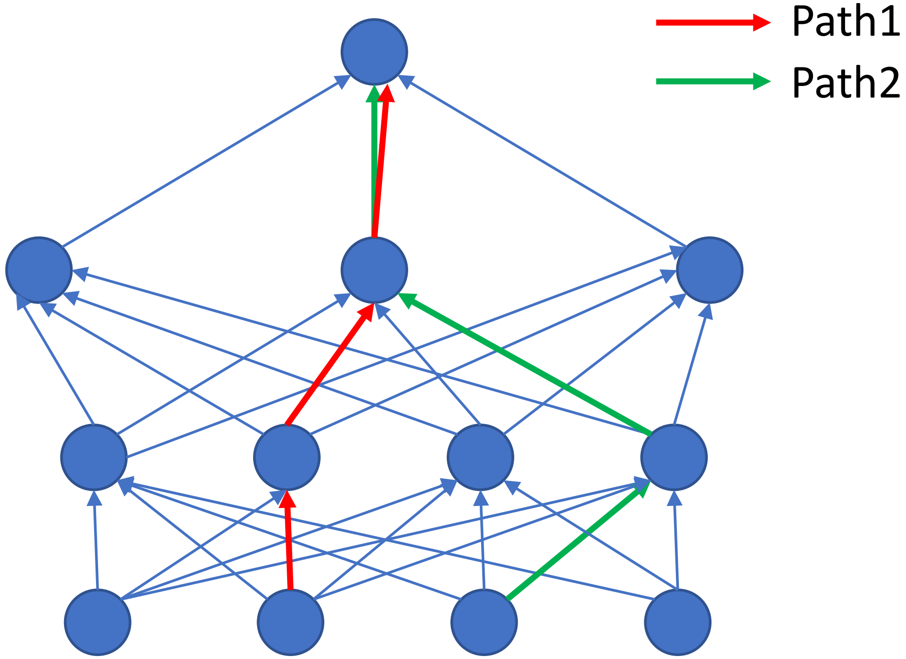
Related Works

- Independence of Activations / Path / Weights

Add noise to make
them independent

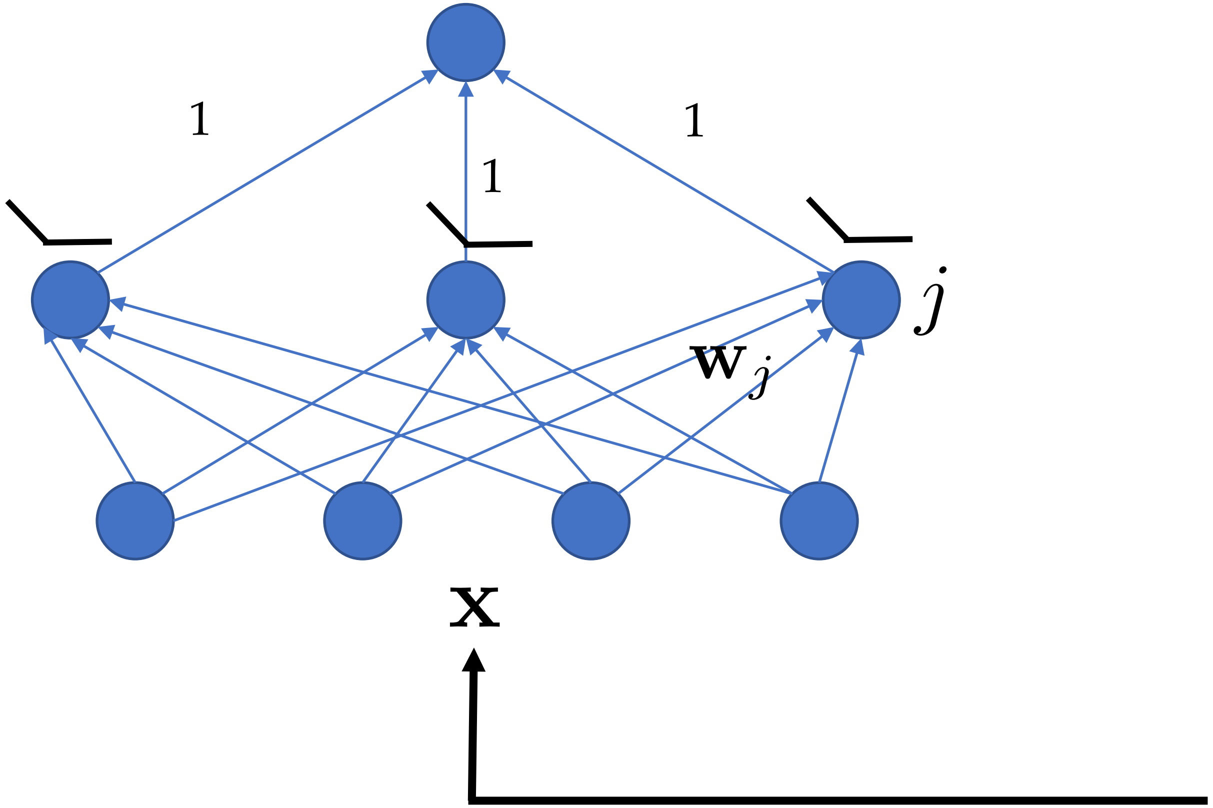


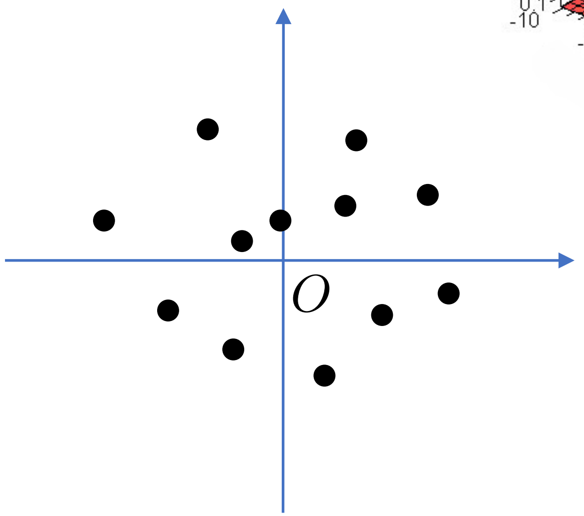
No bad local minima: Data independent training
error guarantees for multilayer neural networks,
D. Soudry and Y. Carmon

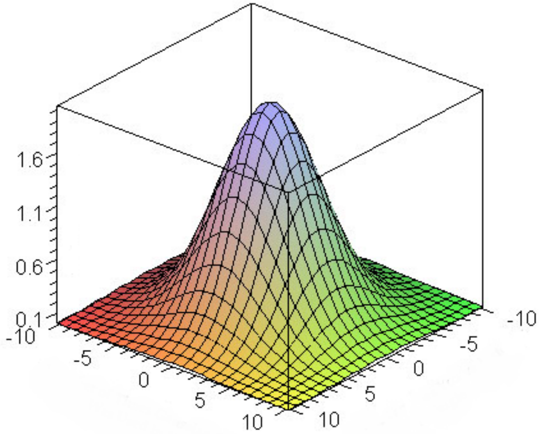


The Loss Surfaces of Multilayer Networks,
A. Choromanska et al, AISTATS 2015

Gaussian Assumption

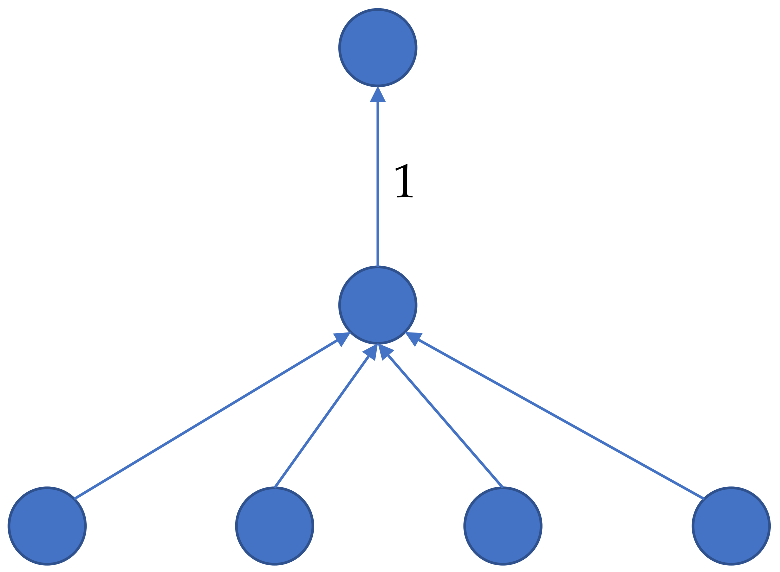






$$\mathbf{x} \sim N(0, I)$$

An Analytic Formula for a single hidden node



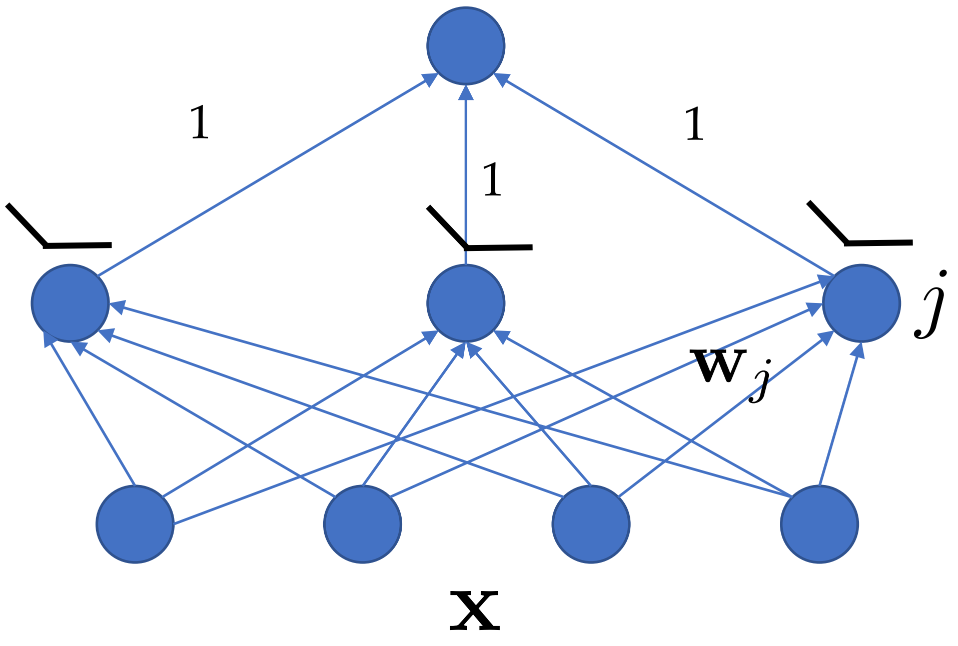
Close-form Population Gradient:

$$\mathbb{E}\left[\nabla_{\mathbf{w}} J\right] = \boxed{\frac{N}{2}(\mathbf{w} - \mathbf{w}^*)} + \boxed{\frac{N}{2\pi}\left(\theta \mathbf{w}^* - \frac{\|\mathbf{w}^*\|}{\|\mathbf{w}\|}\sin\theta \mathbf{w}\right)}$$

Linear component.

Nonlinear component due to ReLU gating

An Analytic Formula



Close-form Population Gradient:

$$\mathbb{E}[\nabla_{\mathbf{w}_j} J] = \sum_{j'=1}^{K} \mathbb{E}[F(\mathbf{e}_j, \mathbf{w}'_j)] - \sum_{j'=1}^{K} \mathbb{E}[F(\mathbf{e}_j, \mathbf{w}^*_{j'})]$$

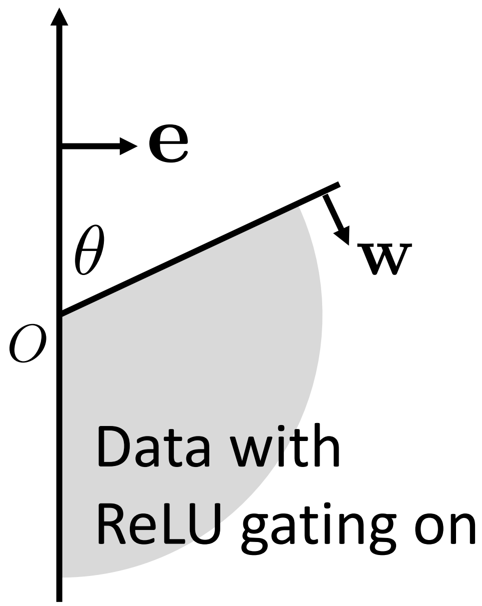
$$F(\mathbf{e}, \mathbf{w}) = X^T D(\mathbf{e}) D(\mathbf{w}) X \mathbf{w}$$

$$\mathbb{E}[F(\mathbf{e}, \mathbf{w})] = \frac{N}{2\pi} [(\pi - \theta)\mathbf{w} + \|\mathbf{w}\| \sin\theta \mathbf{e}]$$

$$\mathbf{e}_j = \mathbf{w}_j / \|\mathbf{w}_j\|$$

Derivation

$$F(\mathbf{e}, \mathbf{w}) = X^T D(\mathbf{e}) D(\mathbf{w}) X \mathbf{w}$$



Alternative Derivation

$$\mathbb{E}_{\mathbf{x}} J(\mathbf{x}; \mathbf{w}) = \sum_{j,k} \left[\phi_1(\mathbf{w}_j, \mathbf{w}_k) - 2\phi_1(\mathbf{w}_j^*, \mathbf{w}_k) + \phi_1(\mathbf{w}_j^*, \mathbf{w}_k^*)\right]^2$$

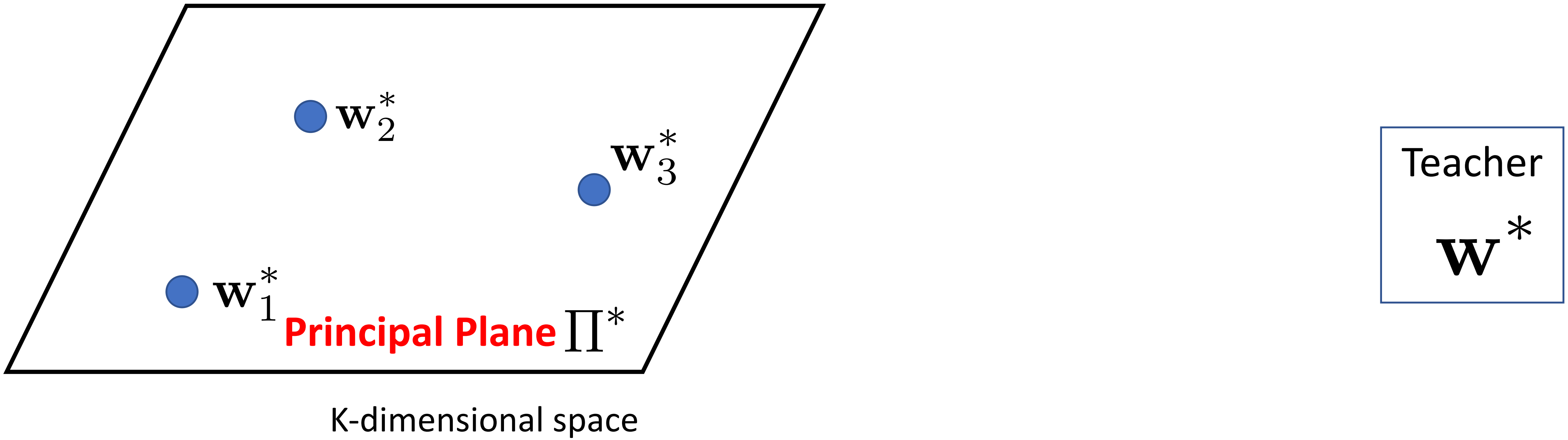
$$\phi_1(\mathbf{w}_j, \mathbf{w}_k) = \mathbb{E}_{\mathbf{x}}\left[\sigma(\mathbf{w}_j^T \mathbf{x}) \sigma(\mathbf{w}_k^T \mathbf{x})\right]$$

Has close form

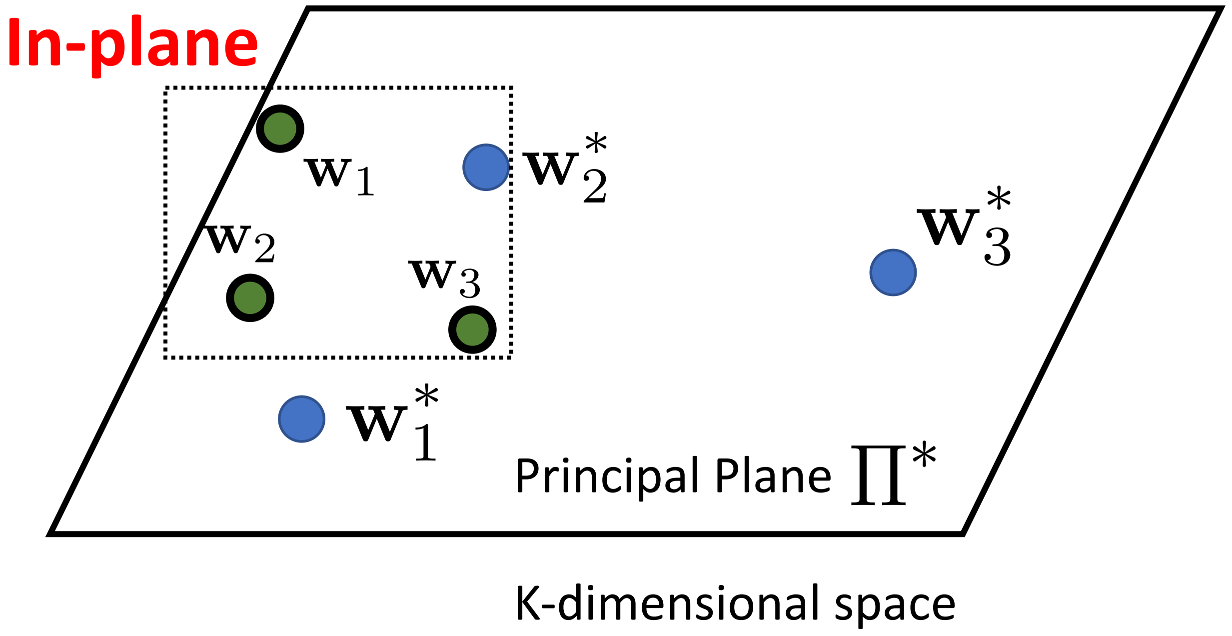
Globally Optimal Gradient Descent for a ConvNet with Gaussian Inputs,
A. Brutzkus and A. Globerson, ICML 2017

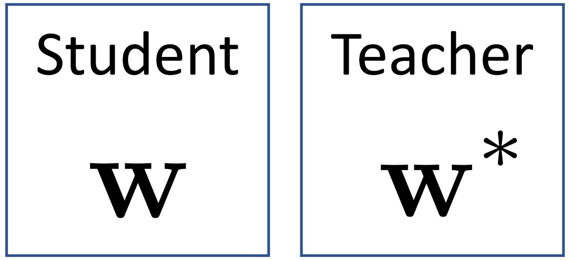
Kernel Methods for Deep Learning, Y. Cho and L. K. Saul, NIPS 2009

Critical Point Analysis



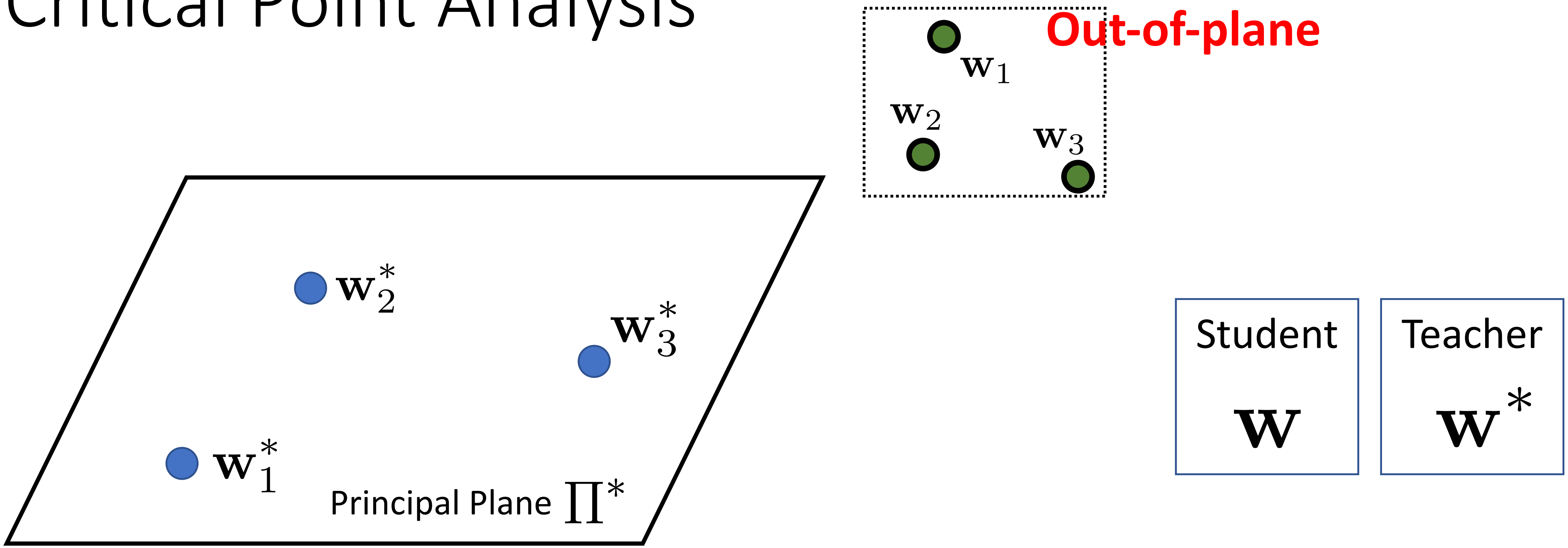
Critical Point Analysis





In-plane: All $\mathbf{w}_j \in \prod^*$

Critical Point Analysis

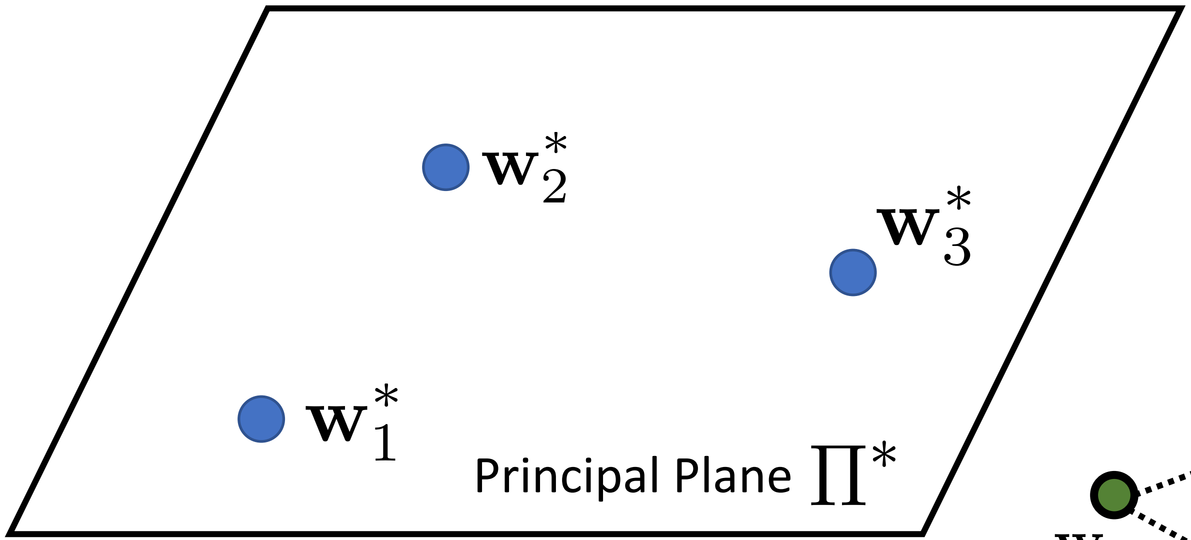


K-dimensional space

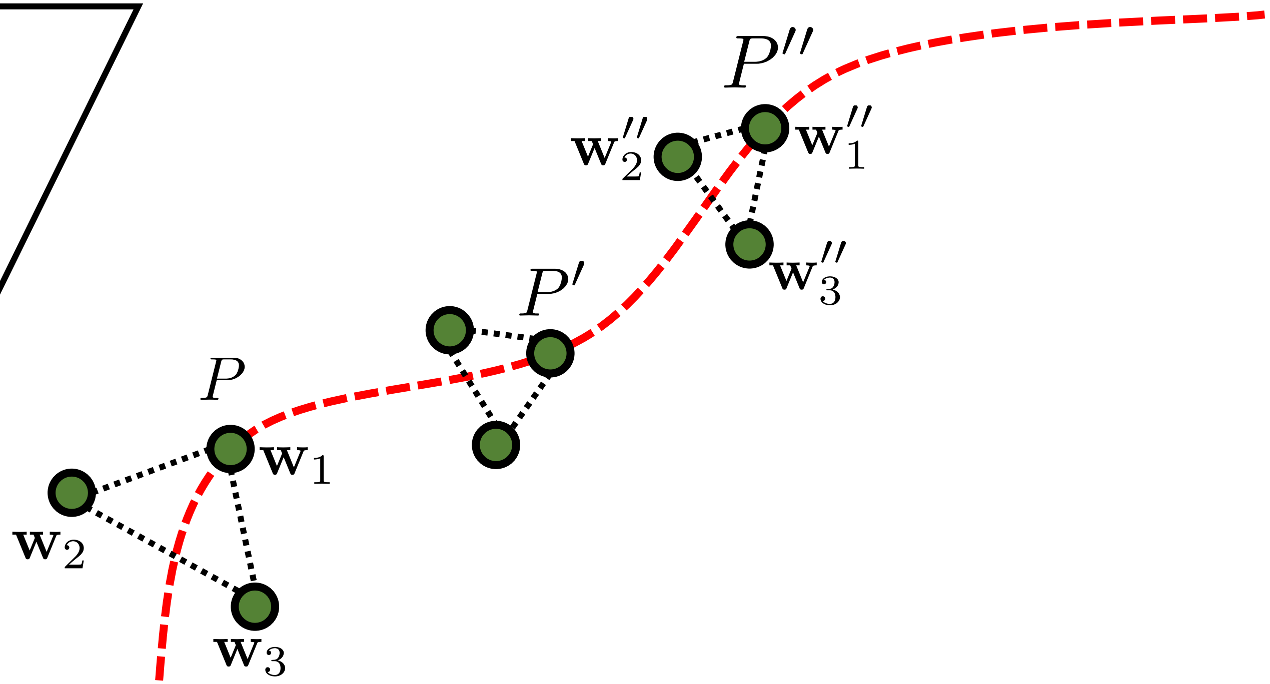
Out-of-plane: Any $\mathbf{w}_j \in \Pi^*$

Critical Point Analysis (Out of plane)

If P is critical point, so does P’ and P’’



K-dimensional space

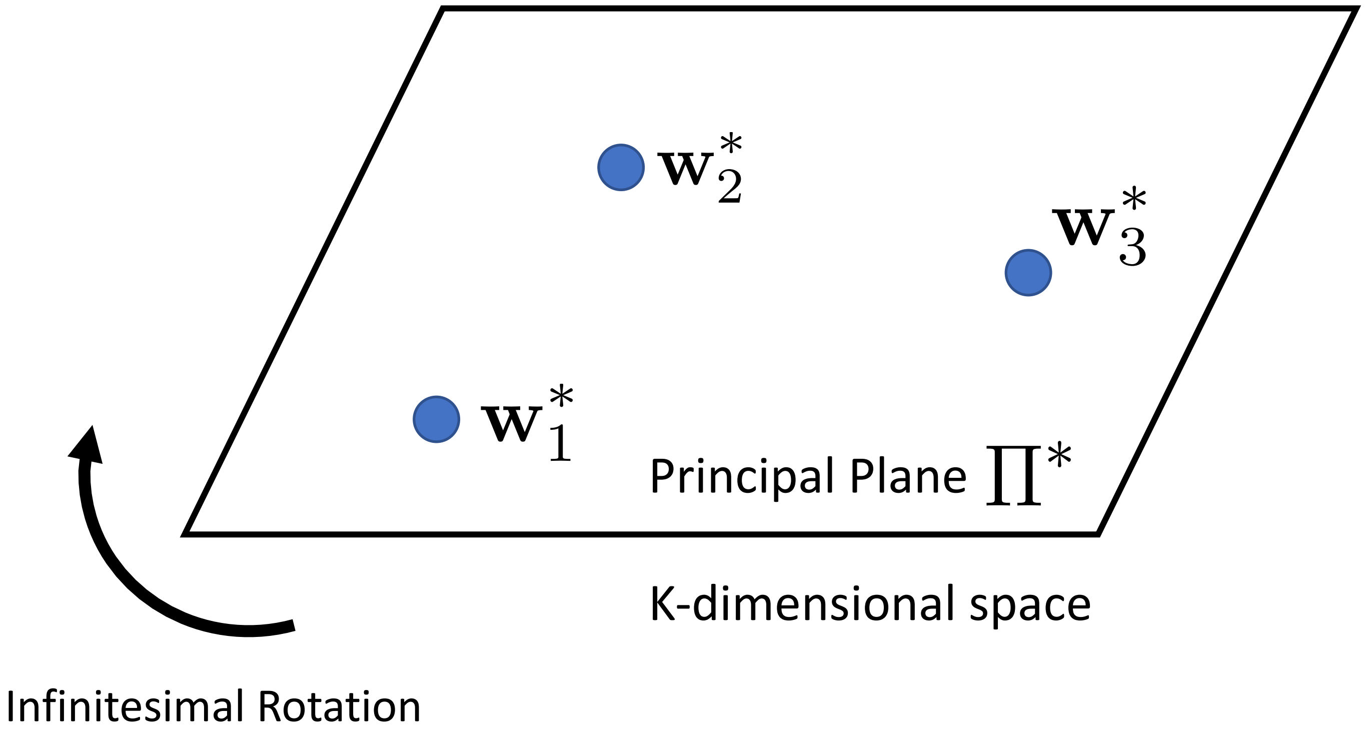


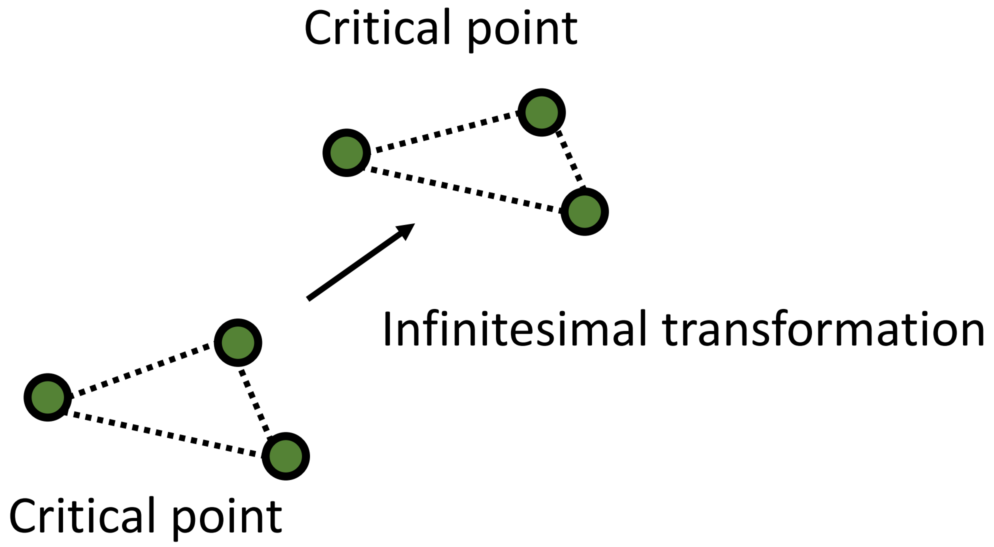
Theorem 2 *If $d \geq K + 2$, then out-of-plane critical points (solutions of Eqn. 9) are non-isolated and lie in a manifold.*

Why we have such structure?

Critical Point Analysis (Out of plane)

How to prove?

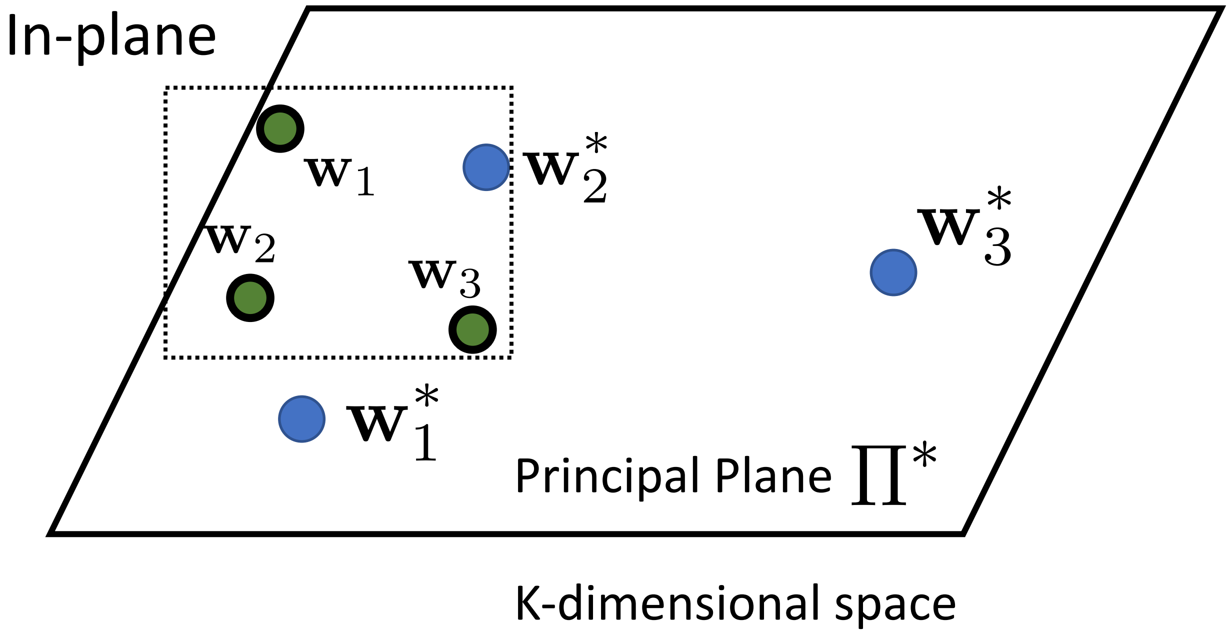


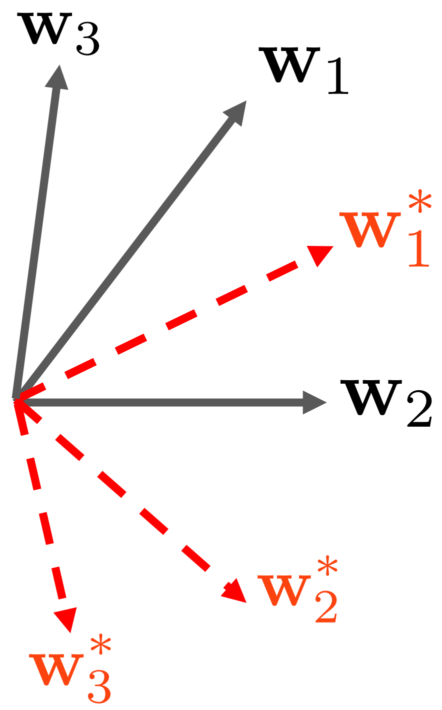


Principle Plane ***invariant*** under rotation

➔ Invariant Objective

Critical Point Analysis (In-plane)

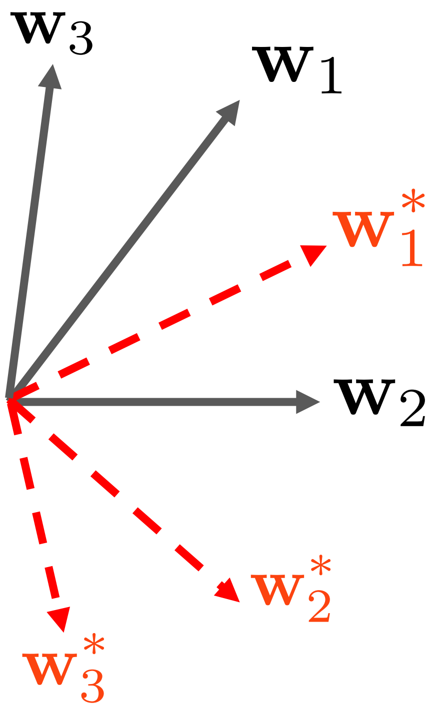




Is this a critical point?

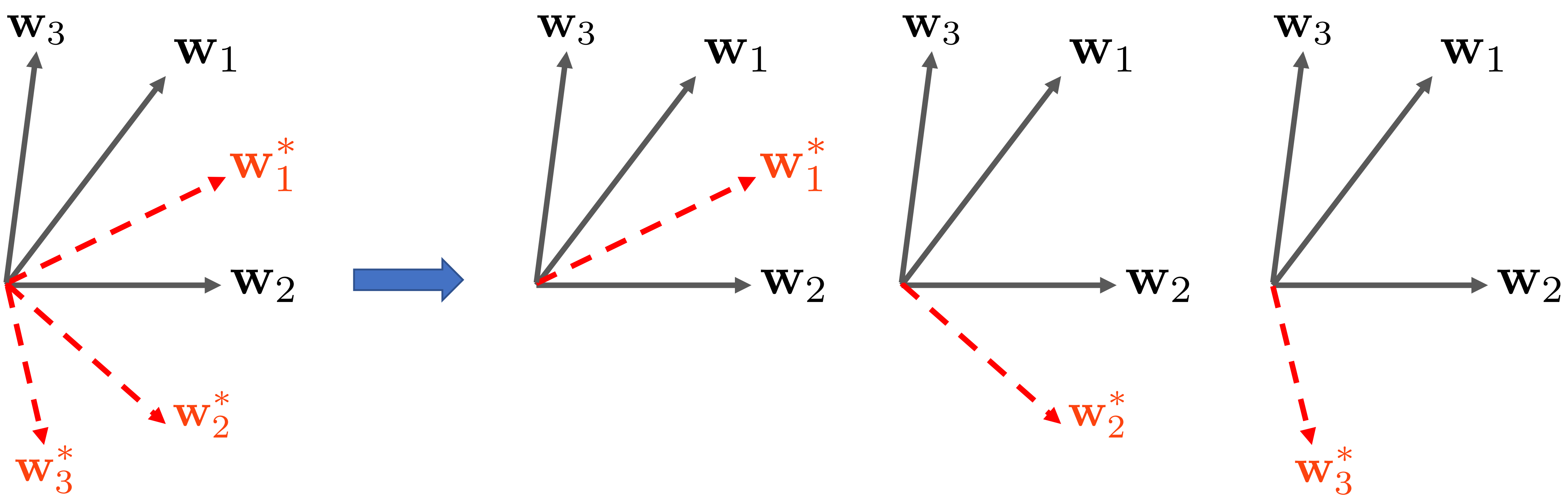
Critical Point Analysis (In-plane)

Decomposition:



Critical Point Analysis (In-plane)

Decomposition:

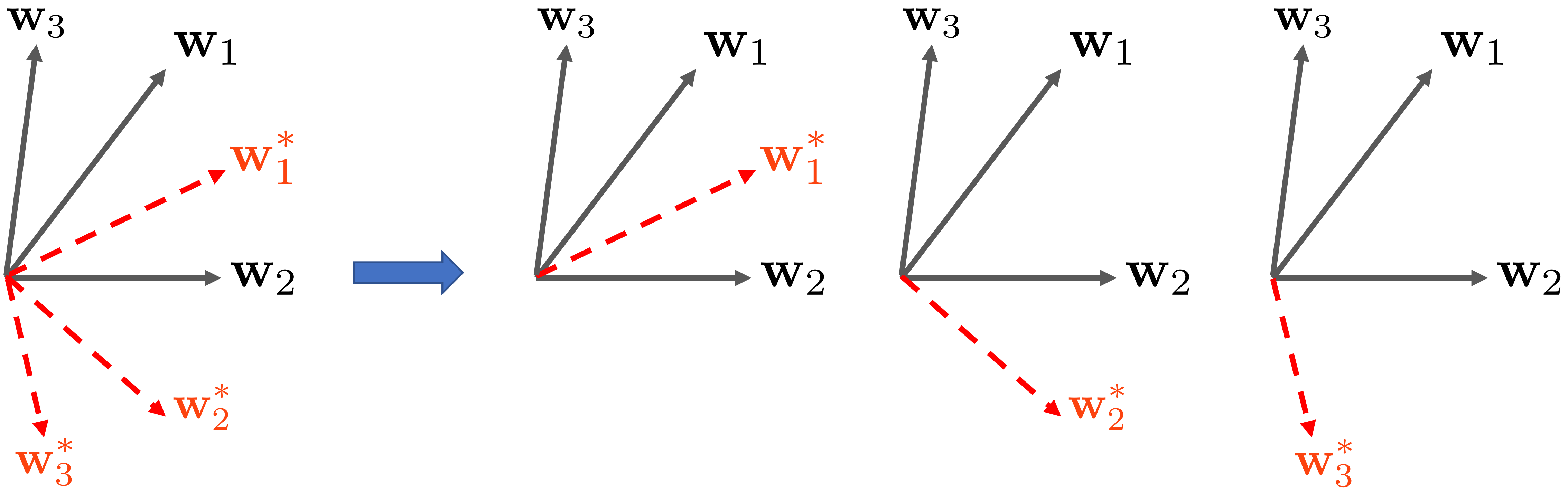


Design a quantity L_1 L_2 L_3

so that each quantity depends on only one vector $\mathbf{w}_j^*$

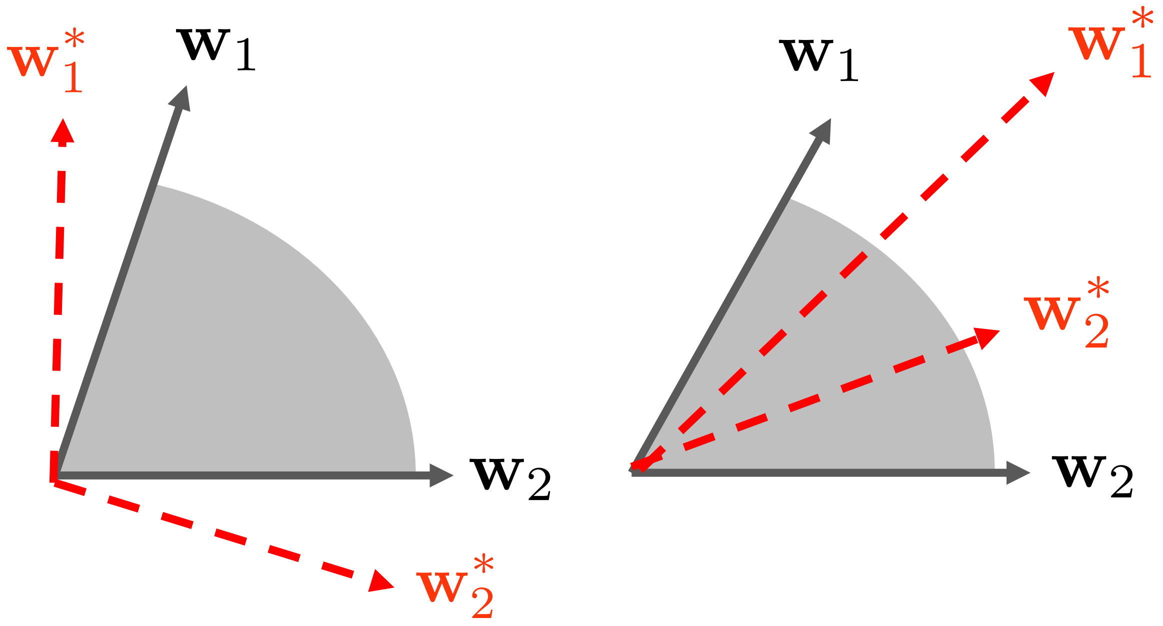
Critical Point Analysis (In-plane)

Decomposition:

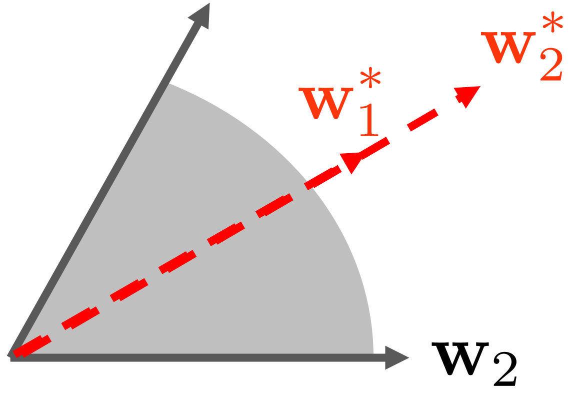


Theorem 3 *If* $\bar{\mathbf{w}}^* \neq 0$, *and for a given parameter* $\mathbf{w}$, $L_{jj'}(\{\theta_l^{*k}\}, \Theta) > 0$ *(or* < 0*) for all* $1 \leq k \leq K$, *then* $\mathbf{w}$ *cannot be a critical point.*

Critical Point Analysis (K=2)

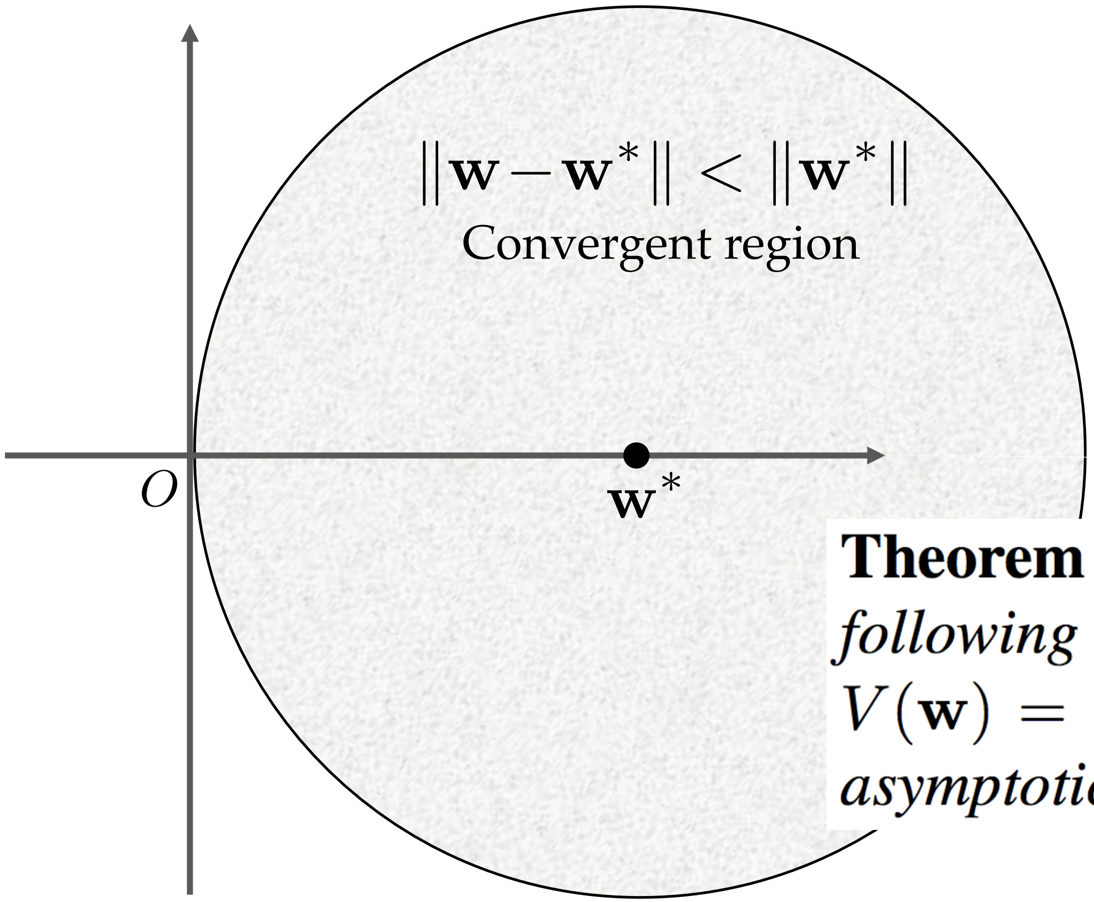


Not a critical point



Saddle Points

Convergence Analysis (single node)



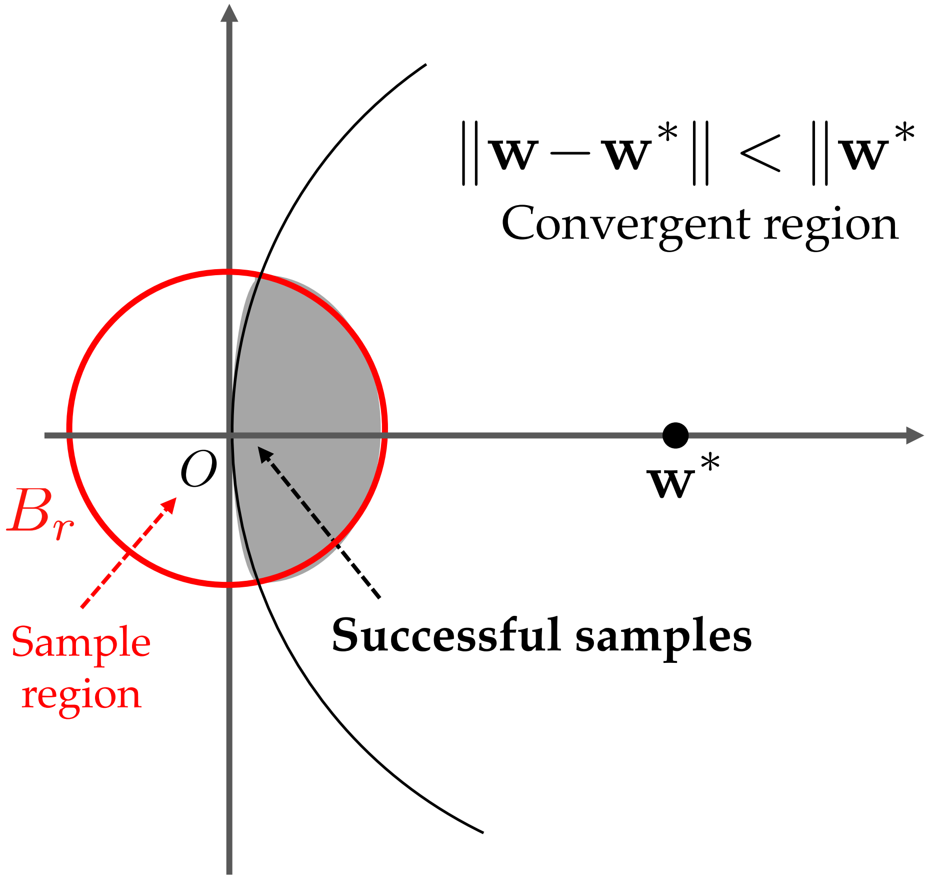
$$\mathbb{E}_{\mathbf{x}} J(\mathbf{x};\mathbf{w}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[\|\sigma(\mathbf{x};\mathbf{w}^*) - \sigma(\mathbf{x};\mathbf{w})\|^2\right]$$

Lyapunov function $V(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}-\mathbf{w}^*\|^2$

Theorem 5 *When* $\mathbf{w}^0 \in \Omega = \{\mathbf{w} : \|\mathbf{w}-\mathbf{w}^*\| < \|\mathbf{w}^*\|\}$, *following the dynamics of Eqn. 15, the Lyapunov function* $V(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}-\mathbf{w}^*\|^2$ *has* $\mathrm{d}V/\mathrm{d}t < 0$ *and the system is asymptotically stable and thus* $\mathbf{w}^t \to \mathbf{w}^*$ *when* $t \to +\infty$.

Convergence Analysis (single node)

Sampling Strategy

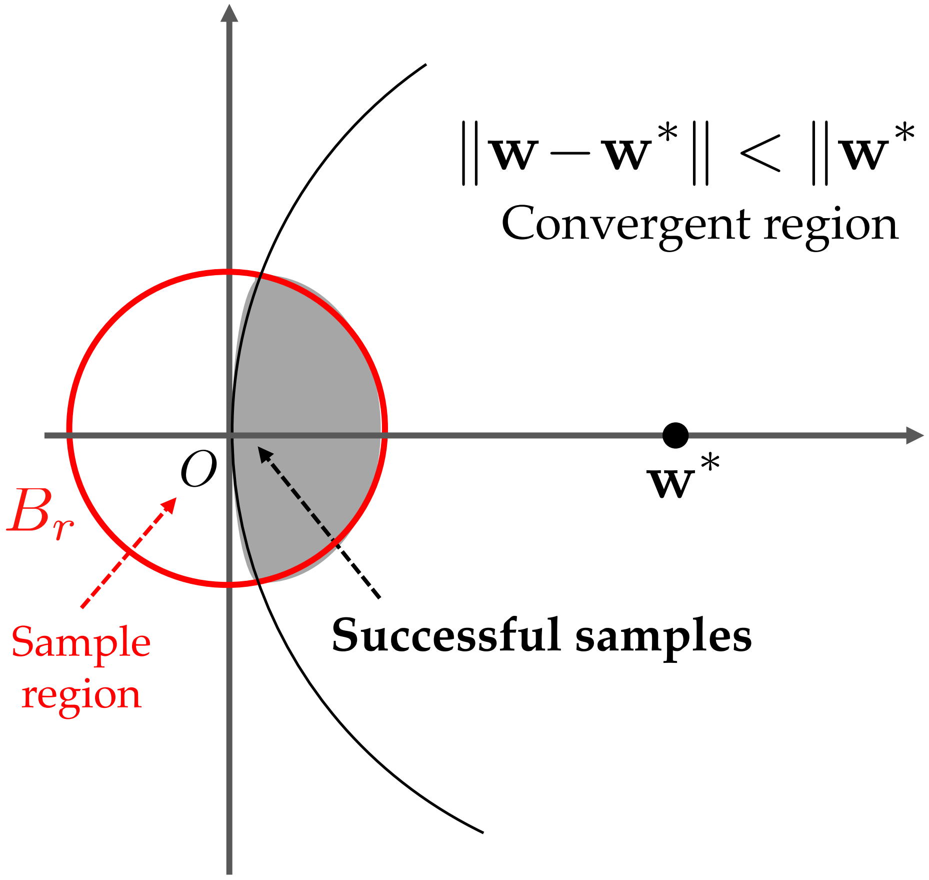


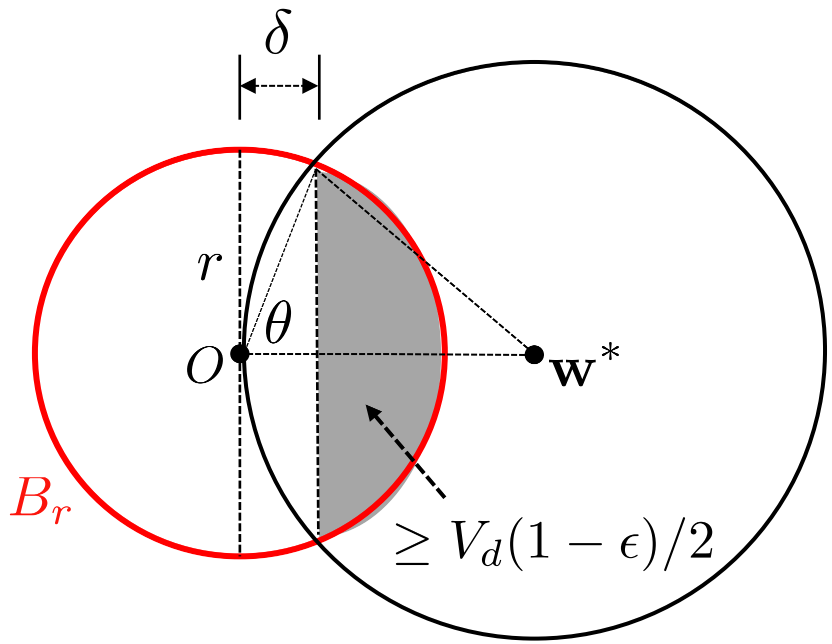
B_r Large → Negligible probability

B_r Small → Locally hyperplane: p ~1/2

Convergence Analysis (single node)

Sampling Strategy:





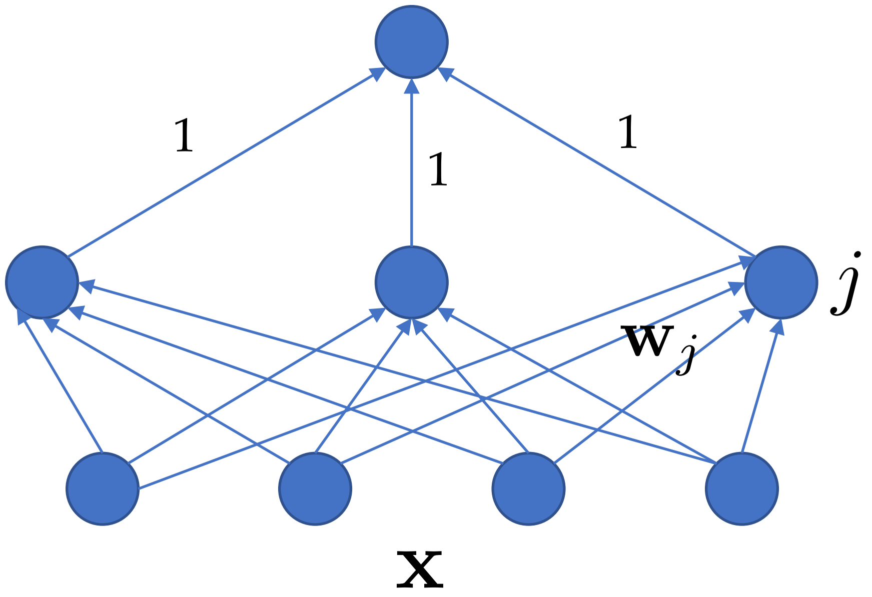
Theorem 6 *The dynamics in Eqn. 6 converges to* $\mathbf{w}^*$ *with probability at least* $(1-\epsilon)/2$*, if the initial value* $\mathbf{w}^0$ *is sampled uniformly from* $B_r = \{\mathbf{w} : \|\mathbf{w}\| \leq r\}$ *with* $r \leq \epsilon\sqrt{\frac{2\pi}{d+1}}\|\mathbf{w}^*\|$.

$r \sim O(1/\sqrt{d})$

Xavier/Kaiming initialization

Convergence Analysis (multiple nodes)

$$g(\mathbf{x};\mathbf{w}) = \sum_{j=1}^{K} \sigma(\mathbf{w}_j^T \mathbf{x})$$



$$\sigma(x) = \max(x, 0)$$

Condition 1: Symmetric weights: $\mathcal{G} = \{P_j\}$

$$\mathbf{w}_j = P_j \mathbf{w}_1$$

$$\mathbf{w}_j^* = P_j \mathbf{w}_1^*$$

Condition 2: Orthonormal ground truth weights:

$$\mathbf{w}_j^{*T} \mathbf{w}_k^* = \delta_{jk}$$

Condition 3: Initial $\mathbf{w}^0$ is in-plane

One special case of Cond1 + Cond2:
Non-overlapping shared weights

Convergence Analysis (multiple nodes)

Reduce to 2D system (x, y) $\mathbf{w}_1 = [x, y, \ldots, y]^T$ under the basis of $\{\mathbf{w}_j^*\}_{j=1}^K$

$$-\frac{2\pi}{N}\mathbb{E}\begin{bmatrix}\nabla_x J \\ \nabla_y J\end{bmatrix} = -\left\{[(\pi-\phi)(x-1+(K-1)y)]\begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}\theta \\ \phi^*-\phi\end{bmatrix} + \phi\begin{bmatrix}x-1\\y\end{bmatrix}\right\}$$
$$+ \quad [(K-1)(\alpha\sin\phi^* - \sin\phi) + \alpha\sin\theta]\begin{bmatrix}x\\y\end{bmatrix}$$

$$\alpha = (x^2+(K-1)y^2)^{-1/2}, \quad \cos\theta = \alpha x, \quad \cos\phi^* = \alpha y, \quad \cos\phi = \alpha^2(2xy+(K-2)y^2)$$

Theorem 7

(1) When the student parameters is initialized to be $[x^0, y^0, \ldots, y^0]$ under the basis of $\mathbf{w}^*$, where $(x^0, y^0) \in \Omega = \{x \in (0,1], y \in [0,1], x > y\}$, then the dynamics (Eqn. 64) converges to teacher's parameters $\{\mathbf{w}_j^*\}$ (or $(x,y) = (1,0)$);

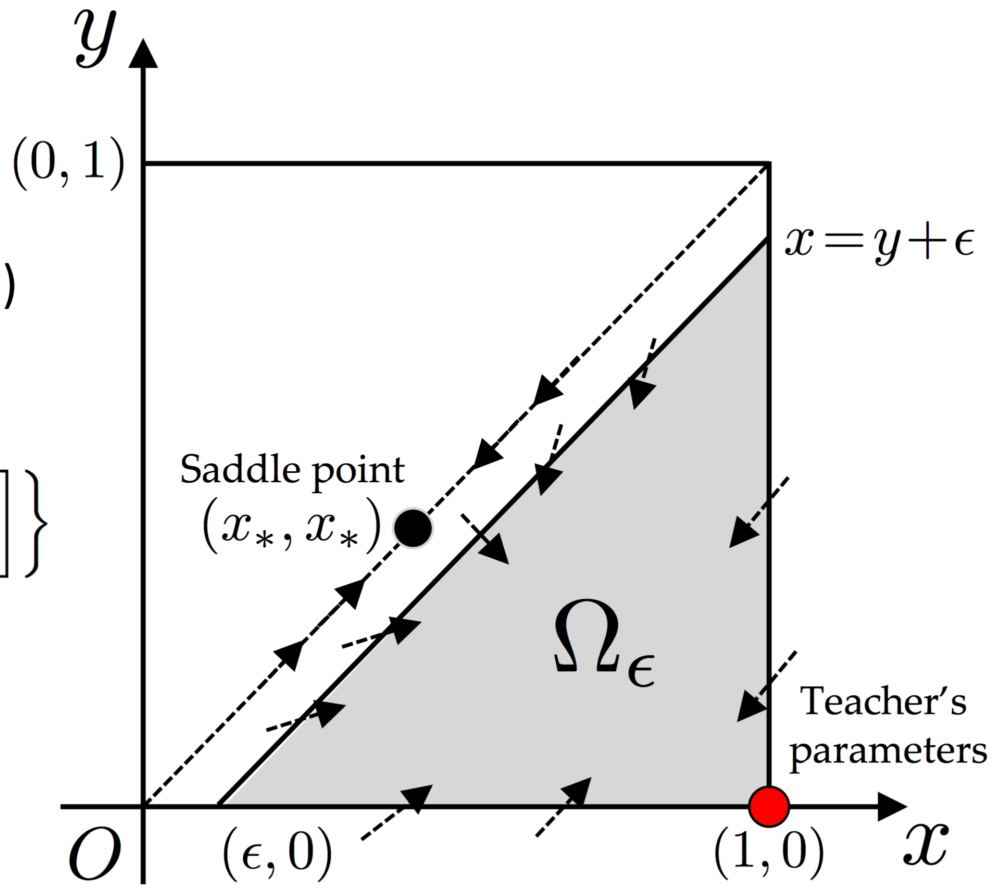
(2) when $x^0 = y^0 \in (0,1]$, then it converges to a saddle point $x = y = \frac{1}{\pi K}(\sqrt{K-1} - \arccos(1/\sqrt{K}) + \pi)$.

2D dynamics

How to prove?

1. Prove the region is convergent
2. Prove within the region, there is no critical point except for (1, 0)
   (via reparametrization)

$$-\frac{2\pi}{N}\mathbb{E}\begin{bmatrix}\nabla_x J\\ \nabla_y J\end{bmatrix} = -\left\{[(\pi-\phi)(x-1+(K-1)y)]\begin{bmatrix}1\\1\end{bmatrix}+\begin{bmatrix}\theta\\ \phi^*-\phi\end{bmatrix}+\phi\begin{bmatrix}x-1\\ y\end{bmatrix}\right\}$$
$$+\quad [(K-1)(\alpha\sin\phi^*-\sin\phi)+\alpha\sin\theta]\begin{bmatrix}x\\ y\end{bmatrix}$$

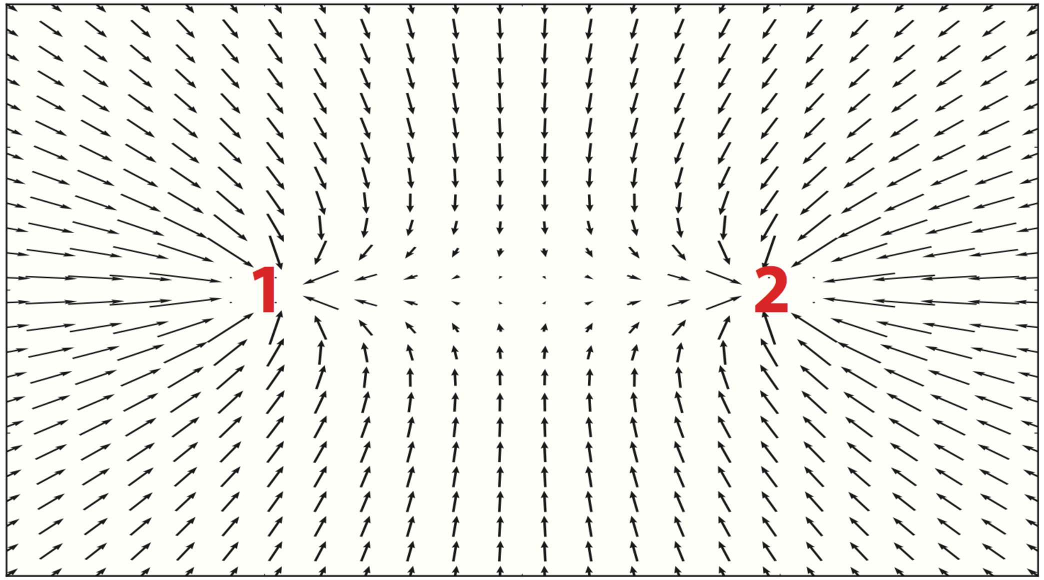


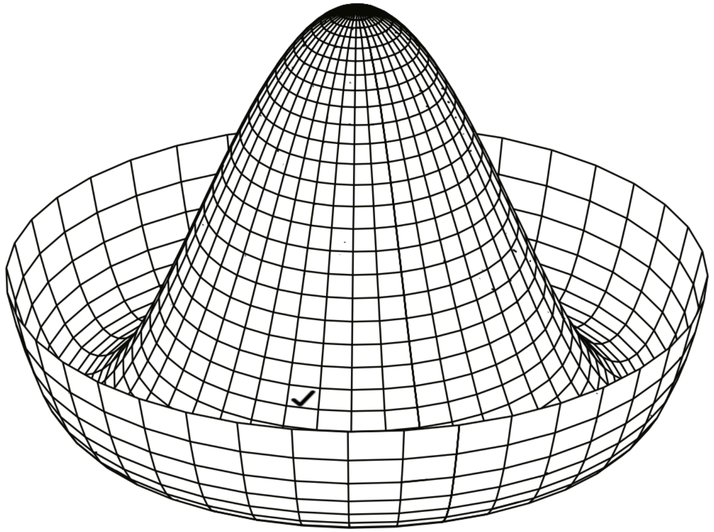
Spontaneous Symmetry Breaking

$\mathbf{w}_1^0 = [x, x, \ldots, x]^\intercal$ Converges to saddle

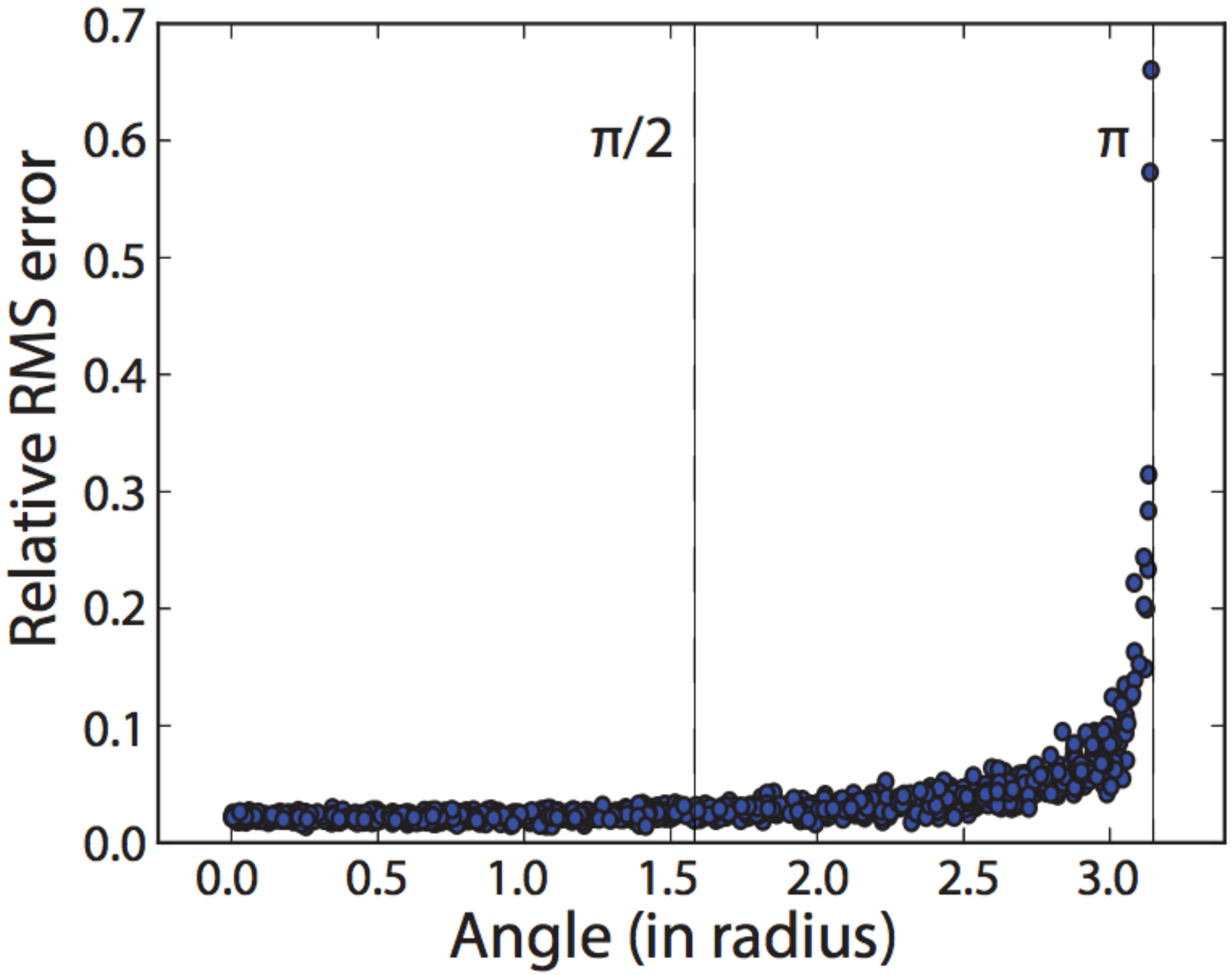
$\mathbf{w}_1^0 = [x + \epsilon, x, \ldots, x]^\intercal$ Converges to $\mathbf{w}^*$

$\mathbf{w}_1^0 = [x, x + \epsilon, \ldots, x]^\intercal$ Converges to permutation of $\mathbf{w}^*$



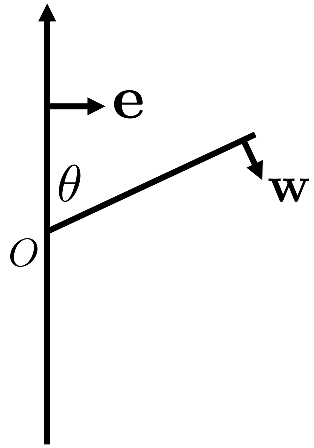


Numerical Experiments



$$F(\mathbf{e}, \mathbf{w}) = X^\intercal D(\mathbf{e}) D(\mathbf{w}) X \mathbf{w}$$

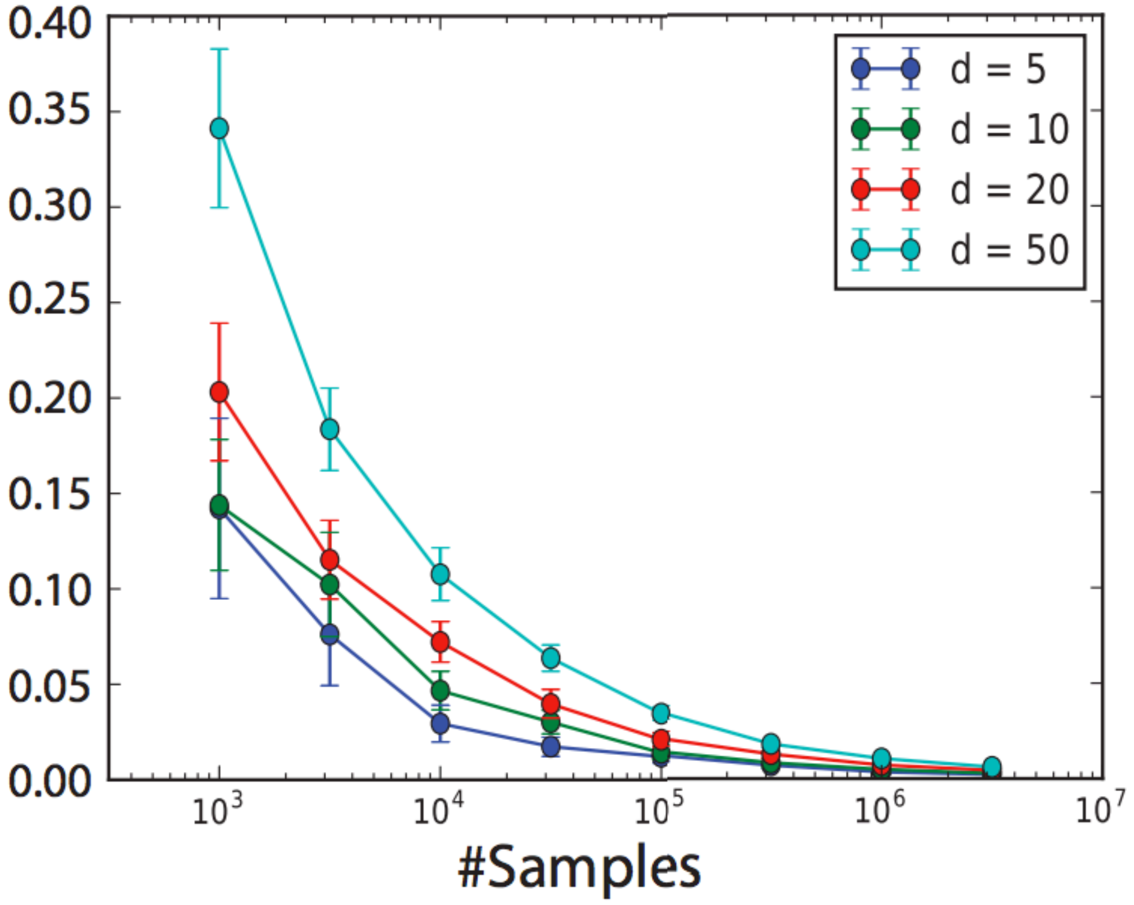
$$\mathbb{E}\left[F(\mathbf{e}, \mathbf{w})\right] = \frac{N}{2\pi}\left[(\pi - \theta)\mathbf{w} + \|\mathbf{w}\| \sin\theta \mathbf{e}\right]$$



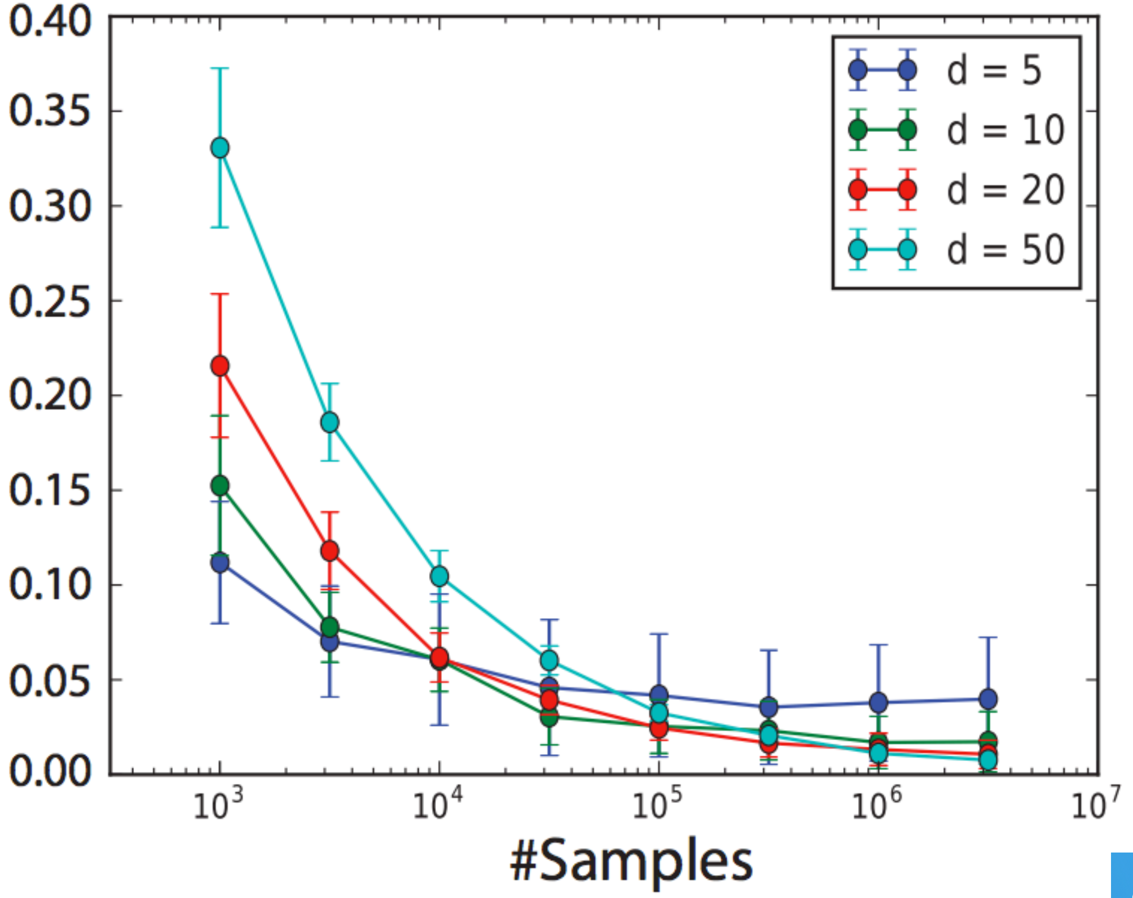
Numerical Experiments

$$\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right] = \frac{N}{2\pi}\left[(\pi-\theta)\mathbf{w} + \|\mathbf{w}\| \sin\theta \mathbf{e}\right]$$

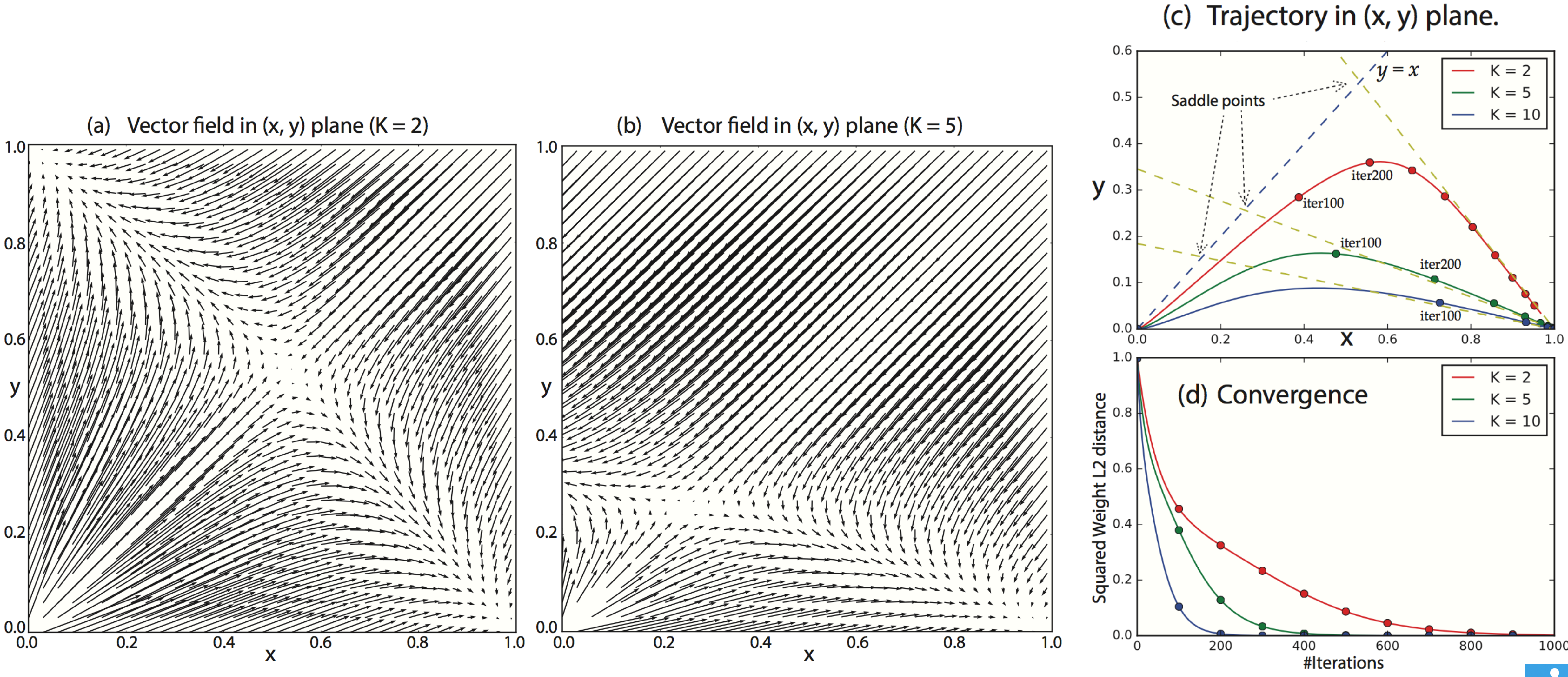
Gaussian Distribution



Uniform Distribution



Numerical Experiments (2D dynamics)



Summary

- An analytic form of population gradient for 2-layered ReLU network with Gaussian input
- Critical Point Analysis
  - Out-of-plane
  - In-plane
- Convergence Analysis
  - Single node
  - Multiple node (special symmetric case)

Future Work

- Non-Gaussian Case?

Isotropic: $\mathbb{E}_{\mathbf{x}}\left[F(\mathbf{e}, \mathbf{w})\right] = A(\theta)\mathbf{w} + \|\mathbf{w}\|B(\theta)\mathbf{e}$

More general distributions?

- Multilayered nonlinear network

Proposition 2 *Denote $[c]$ as all nodes in layer c. Denote $\mathbf{u}_j^*$ and $\mathbf{u}_j$ as the output of node j at layer c of the teacher and student network, then the gradient of the parameters $\mathbf{w}_j$ immediate under node $j \in [c]$ is:*

$$\nabla_{\mathbf{w}_j} J = X_c^\intercal D_j Q_j \sum_{j' \in [c]} \left(Q_{j'}\mathbf{u}_{j'} - Q_{j'}^*\mathbf{u}_{j'}^*\right) \qquad (19)$$

where X_c is the data fed into node j, Q_j and Q_j^* are N-by-N diagonal matrices. For any node $k \in [c+1]$, $Q_k = \sum_{j\in[c]} w_{jk} D_j Q_j$ and similarly for Q_k^*.

Thanks!

Poster #85