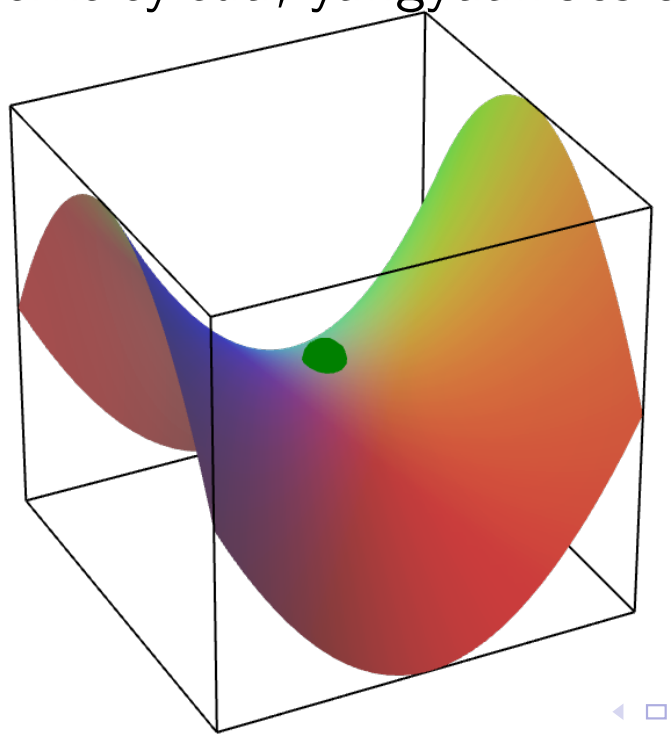
Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition

Rong Ge, Furong Huang, Chi Jin, Yang Yuan

rongge@microsoft.com, furongh@uci.edu
chijin@cs.berkeley.edu, yangyuan@cs.cornell.edu



Outline

Outline

Stochastic Gradient Descent

Stochastic optimization problem: $w = \arg\min_{w \in \mathbb{R}^d} f(w)$

$$\boxed{w_{t+1} = w_t - \eta SG(w_t)}$$

where $\mathbb{E}[SG(w)] = \nabla f(w)$, $\|SG(w) - \nabla f(w)\| \leq Q$.

Stochastic Gradient Descent

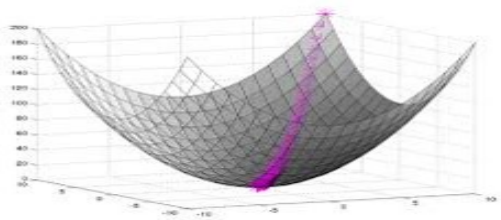
Stochastic optimization problem: $w = \arg\min_{w \in \mathbb{R}^d} f(w)$

$$\boxed{w_{t+1} = w_t - \eta SG(w_t)}$$

where $\mathbb{E}[SG(w)] = \nabla f(w)$, $\|SG(w) - \nabla f(w)\| \leq Q$.

Convex function

- Well studied: (Shalev-Shwartz et. al. COLT09') (Rakhlin et. al. ICML12')
- Neural network: backpropagation (non-convex)
- Spectral methods: decomposition (non-convex)



Stochastic Gradient Descent

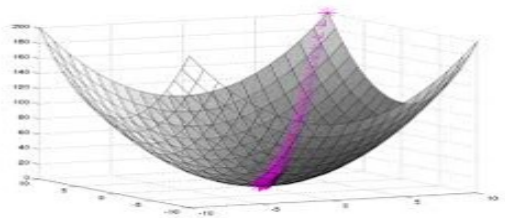
Stochastic optimization problem: $w = \arg\min_{w \in \mathbb{R}^d} f(w)$

$$\boxed{w_{t+1} = w_t - \eta SG(w_t)}$$

where $\mathbb{E}[SG(w)] = \nabla f(w)$, $\|SG(w) - \nabla f(w)\| \leq Q$.

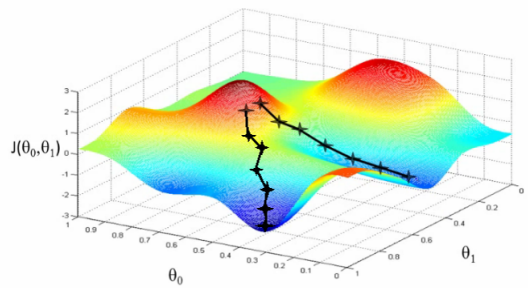
Convex function

- Well studied: (Shalev-Shwartz et. al. COLT09') (Rakhlin et. al. ICML12')
- Neural network: backpropagation (non-convex)
- Spectral methods: decomposition (non-convex)



Non-Convex function

- NP hard
- local minima, finding global minimum is hard
- saddle points, finding local minima is hard



Stochastic Gradient Descent

Stochastic optimization problem: $w = \arg\min_{w \in \mathbb{R}^d} f(w)$

$$\boxed{w_{t+1} = w_t - \eta SG(w_t)}$$

where $\mathbb{E}[SG(w)] = \nabla f(w)$, $\|SG(w) - \nabla f(w)\| \leq Q$.

Stochastic Gradient Descent

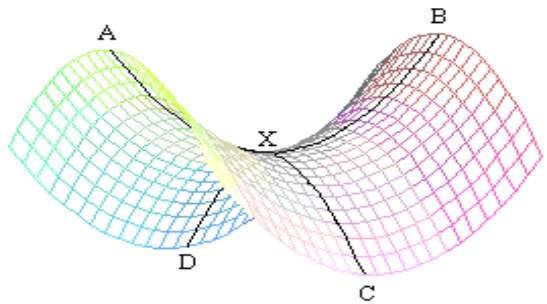
Stochastic optimization problem: $w = \arg\min_{w \in \mathbb{R}^d} f(w)$

$$\boxed{w_{t+1} = w_t - \eta SG(w_t)}$$

where $\mathbb{E}[SG(w)] = \nabla f(w)$, $\|SG(w) - \nabla f(w)\| \leq Q$.

Saddle Point

- saddle points: 0 gradient, not local minima



Question: Is SGD effective in presence of saddle points?

Stochastic Gradient Descent

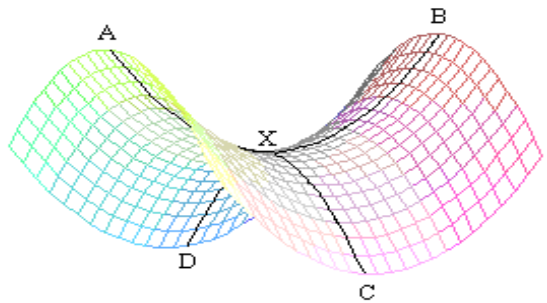
Stochastic optimization problem: $w = \arg\min_{w \in \mathbb{R}^d} f(w)$

$$\boxed{w_{t+1} = w_t - \eta SG(w_t)}$$

where $\mathbb{E}[SG(w)] = \nabla f(w)$, $\|SG(w) - \nabla f(w)\| \leq Q$.

Saddle Point

- saddle points: 0 gradient, not local minima



Question: Is SGD effective in presence of saddle points?

Given a non-convex function with many saddle points, what properties will guarantee stochastic gradient descent to converge to a local minimum efficiently?

Stochastic Gradient Descent

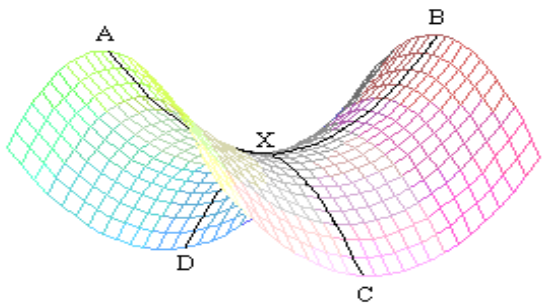
Stochastic optimization problem: $w = \arg\min_{w \in \mathbb{R}^d} f(w)$

$$\boxed{w_{t+1} = w_t - \eta SG(w_t)}$$

where $\mathbb{E}[SG(w)] = \nabla f(w)$, $\|SG(w) - \nabla f(w)\| \leq Q$.

Saddle Point

- saddle points: 0 gradient, not local minima



Question: Is SGD effective in presence of saddle points?

Given a non-convex function with many saddle points, what properties will guarantee stochastic gradient descent to converge to a local minimum efficiently?

Answer: Strict Saddle Property

Outline

Summary of Contribution

- **Identify**: a wide-class of non-convex function called **strict saddle**
  - includes functions with exponentially many local extrema and saddle points.

- **Prove**: SGD converges to a local minimum in polynomial time under *strict saddle*

- **Apply**: Orthogonal tensor decomposition
  - First online first-order algorithm with global convergence guarantee

Outline

Outline

Strict Saddle Functions

Question: What properties guarantee local progress?

Saddle Points

- Stationary Point $\nabla f(w) = 0$: Local maximum/minimum and saddle point
- $\nabla^2 f(w)$ has positive and negative eigenvalues, $\rightarrow w$ saddle point

Definition?

Strict Saddle Functions

Question: What properties guarantee local progress?

Saddle Points

- Stationary Point $\nabla f(w) = 0$: Local maximum/minimum and saddle point
- $\nabla^2 f(w)$ has positive and negative eigenvalues, $\rightarrow w$ saddle point

Definition?

Strict Saddle Functions

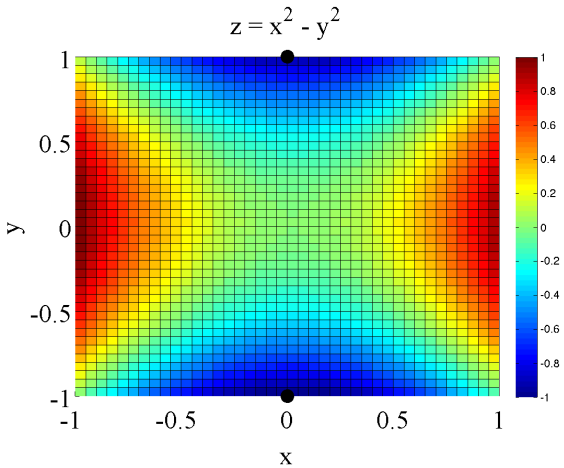
Definition

A twice differentiable $f(w)$ is *strict saddle*, if all its local minima have $\nabla^2 f(w) \succ 0$ and all its other stationary points satisfy $\lambda_{min}(\nabla^2 f(w)) < 0$.

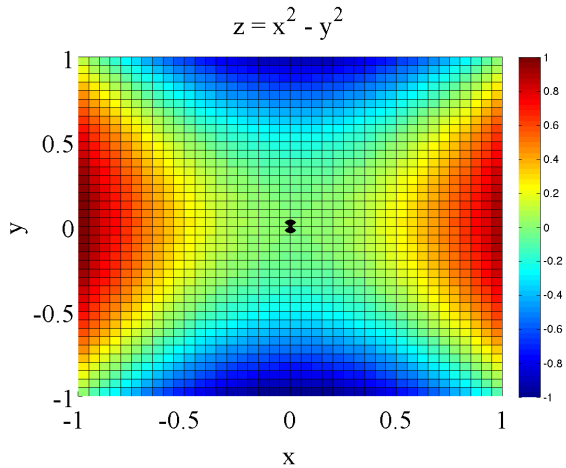
Strict Saddle Functions

Definition

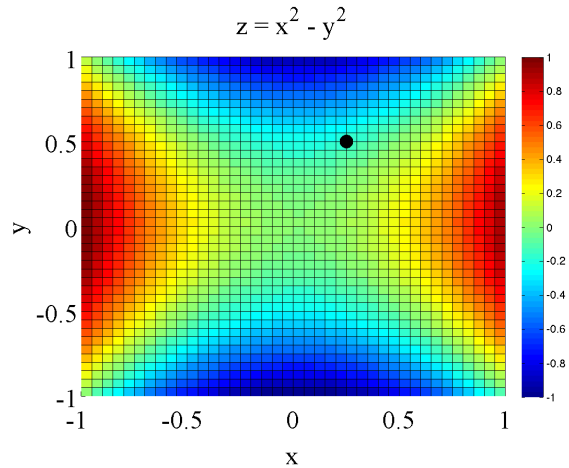
A twice differentiable $f(w)$ is *strict saddle*, if all its local minima have $\nabla^2 f(w) \succ 0$ and all its other stationary points satisfy $\lambda_{min}(\nabla^2 f(w)) < 0$.



Local Min $\|w - w^\star\| \leq \delta$



Saddle $\lambda_{min}(\nabla^2 f(w)) \leq -\gamma$

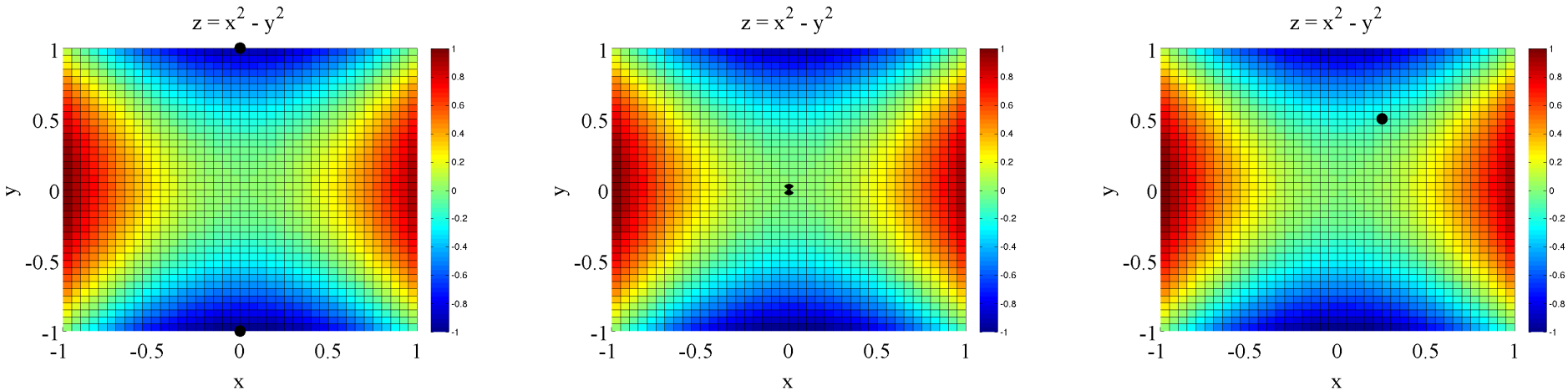


Others $\|\nabla f(w)\| \geq \epsilon$

Strict Saddle Functions

Definition

A twice differentiable $f(w)$ is *strict saddle*, if all its local minima have $\nabla^2 f(w) \succ 0$ and all its other stationary points satisfy $\lambda_{min}(\nabla^2 f(w)) < 0$.



Local Min $\|w - w^\star\| \leq \delta$ Saddle $\lambda_{min}(\nabla^2 f(w)) \leq -\gamma$ Others $\|\nabla f(w)\| \geq \epsilon$

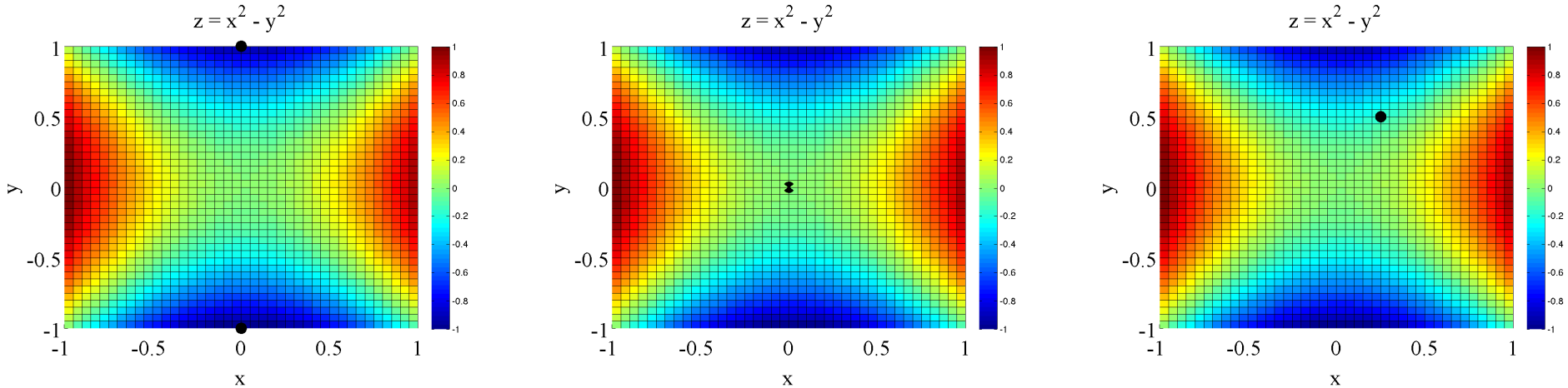
Intuitively guarantees local progress if Hessian info is available

- Second order Taylor expansion of function
- $\nabla^2 f(w)$ negative eigenvalue

Strict Saddle Functions

Definition

A twice differentiable $f(w)$ is *strict saddle*, if all its local minima have $\nabla^2 f(w) \succ 0$ and all its other stationary points satisfy $\lambda_{min}(\nabla^2 f(w)) < 0$.



Local Min $\|w - w^\star\| \leq \delta$ Saddle $\lambda_{min}(\nabla^2 f(w)) \leq -\gamma$ Others $\|\nabla f(w)\| \geq \epsilon$

Intuitively guarantees local progress if Hessian info is available

- Second order Taylor expansion of function
- $\nabla^2 f(w)$ negative eigenvalue

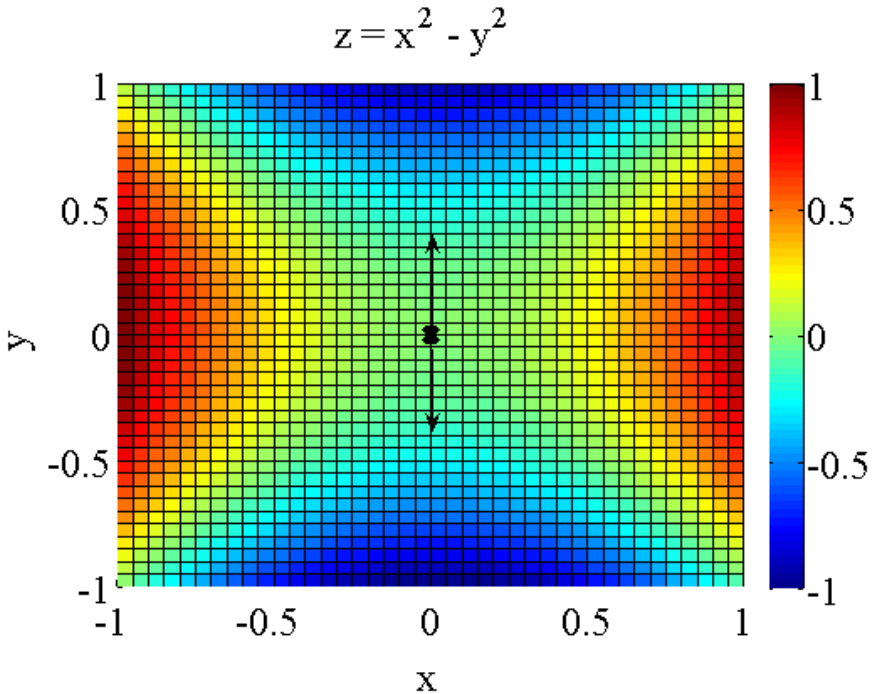
Surprisingly SGD can escape from the saddle points efficiently

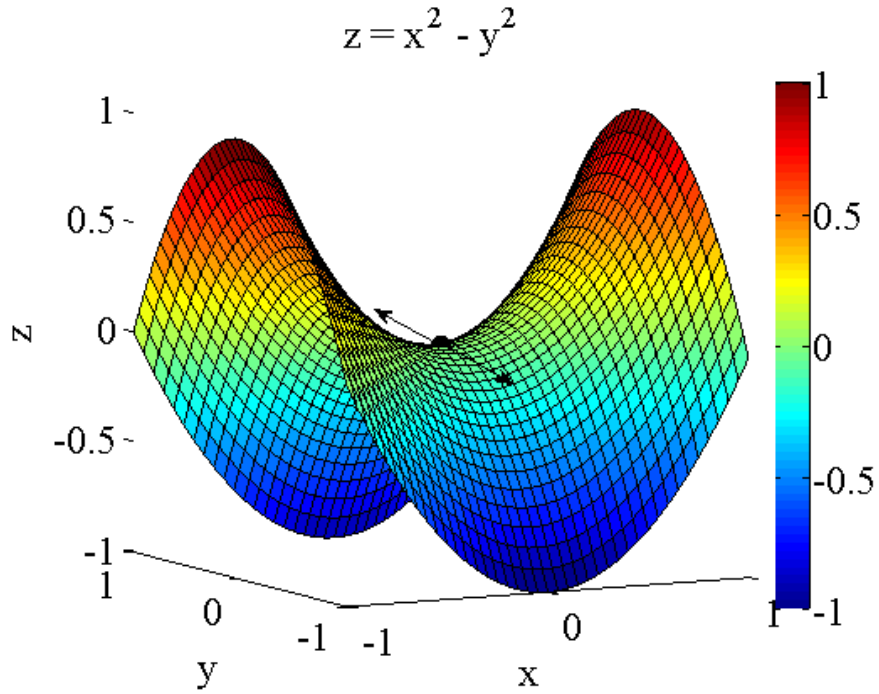
- with only first order gradient information

SGD at Saddle Point

Why SGD works on saddle point?

A toy example $f = x^2 - y^2$:

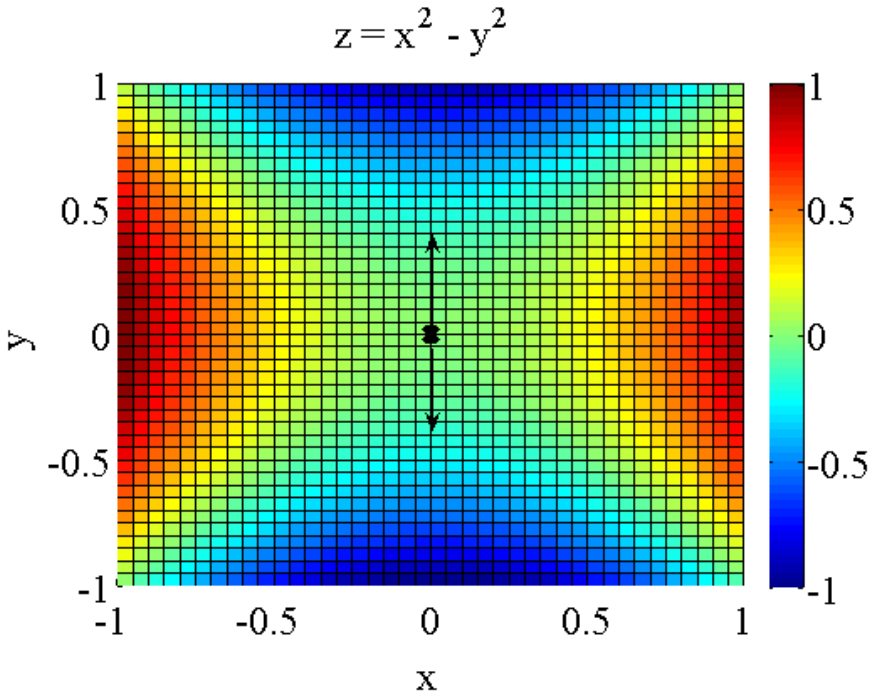


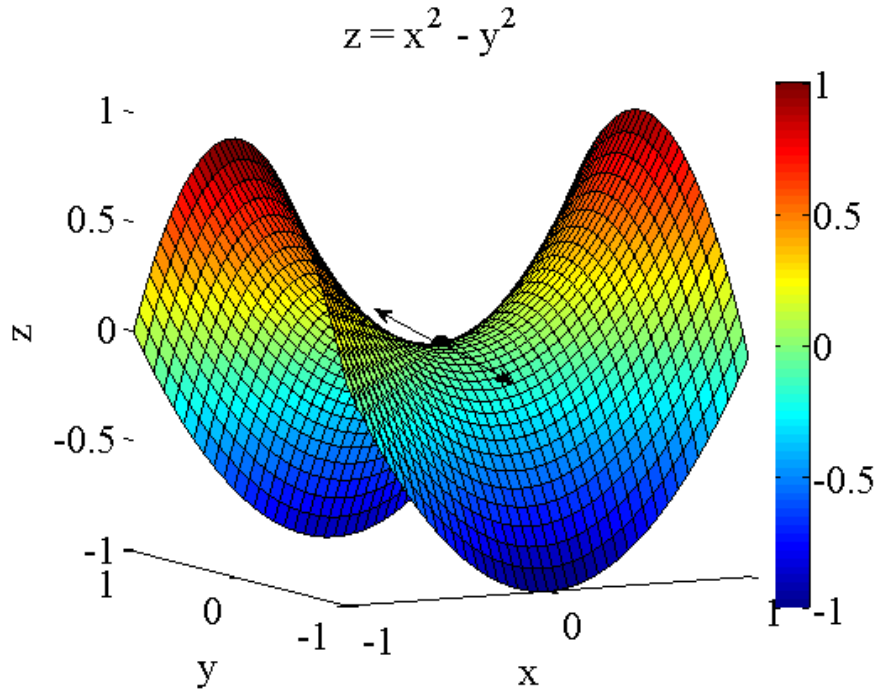


SGD at Saddle Point

Why SGD works on saddle point?

A toy example $f = x^2 - y^2$:





At saddle point $(0, 0)$, SGD movement remains bounded in x direction, but the gradient in y boosts the movement due to SGD. Therefore, we get out of the saddle point.

Outline

SGD for Strict Saddle Functions

Modified Algorithm Noisy Stochastic Gradient Descent (NSGD): At each step, sample noise n uniformly from unit sphere, and

$$w_{t+1} \leftarrow w_t - \eta(SG(w) + n).$$

SGD for Strict Saddle Functions

Modified Algorithm Noisy Stochastic Gradient Descent (NSGD): At each step, sample noise n uniformly from unit sphere, and

$$w_{t+1} \leftarrow w_t - \eta(SG(w) + n).$$

Assume nice smoothness conditions on *strict saddle* function

Outline

Main Result: Convergence Theorem

Theorem: Under assumptions above, for any $\zeta > 0$, there exists a threshold $\eta_{\max}$. So that, as long as the step size $\eta \leq \eta_{\max}$, we have w_T is r-close to some local minimum with probability at least $1 - \zeta$.

Main Result: Convergence Theorem

Theorem: Under assumptions above, for any $\zeta > 0$, there exists a threshold $\eta_{\max}$. So that, as long as the step size $\eta \leq \eta_{\max}$, we have w_T is r-close to some local minimum with probability at least $1 - \zeta$.

$$\eta_{\max} = \tilde{O}(1/\log(1/\zeta)), T = \tilde{O}(\eta^{-2}\log(1/\zeta)), r = \tilde{O}(\sqrt{\eta \log(1/\eta\zeta)})$$

Main Result: Convergence Theorem

Theorem: Under assumptions above, for any $\zeta > 0$, there exists a threshold $\eta_{\max}$. So that, as long as the step size $\eta \leq \eta_{\max}$, we have w_T is r-close to some local minimum with probability at least $1 - \zeta$.

$$\eta_{\max} = \tilde{O}(1/\log(1/\zeta)), T = \tilde{O}(\eta^{-2}\log(1/\zeta)), r = \tilde{O}(\sqrt{\eta \log(1/\eta\zeta)})$$

*__Remarks:__

Main Result: Convergence Theorem

Theorem: Under assumptions above, for any $\zeta > 0$, there exists a threshold $\eta_{\max}$. So that, as long as the step size $\eta \leq \eta_{\max}$, we have w_T is r-close to some local minimum with probability at least $1 - \zeta$.

$$\eta_{\max} = \tilde{O}(1/\log(1/\zeta)), T = \tilde{O}(\eta^{-2}\log(1/\zeta)), r = \tilde{O}(\sqrt{\eta \log(1/\eta\zeta)})$$

*__Remarks:__

- To get $1/\sqrt{t}$ convergence rate, we can use this Theorem to first find a point that is inside the strongly convex region of a local minimum, and then decrease the learning rate by $1/t$.

Main Result: Convergence Theorem

Proof Sketch (informal):

1. [Gradient] When gradient $\|\nabla f(w_t)\|$ is large enough, we have $\mathbb{E}[f(w_{t+1})] \leq f(w_t) - \tilde{\Omega}(\eta^2)$.
2. [Saddle Point] When gradient is small, but $\lambda_{\min}(\nabla^2 f(w_t)) \leq -\gamma$, then we have $\mathbb{E}[f(w_{t+T})] \leq f(w_t) - \tilde{\Omega}(\eta)$ where $T \leq \tilde{O}(1/\eta)$.
3. there is a small probability of being $\tilde{O}(\sqrt{\eta})$-close to a local minimum, $\rightarrow \mathbb{E}[f(w)]$ decrease by at least order η in at most $\tilde{O}(1/\eta)$ iterations.
4. Function value bounded, in $\tilde{O}(1/\eta^2)$ steps with at least constant probability w_t will be $O(\sqrt{\eta})$-close to a local minimum.
5. [Local Minimum] Once w_t is close enough to some local minimum $w^\star$, with probability at least $1 - \zeta/2$, w_{t+i} will still be $\tilde{O}(\sqrt{\eta \log(1/\eta)})$-close to $w^\star$ with all $i < \tilde{O}(\eta^{-2} \log(1/\zeta))$.
6. After $O(\log(1/\zeta))$ epochs of $\tilde{O}(1/\eta^2)$ iterations each, the probability of success will be $1 - \exp(-\log(1/\zeta))$.

Outline

Orthogonal Tensor Decompositon

Motivation

- Latent variable graphical models
- Model parameters are learnt through decomposition of higher order moments

Orthogonal Tensor Decompositon

Motivation

- Latent variable graphical models
- Model parameters are learnt through decomposition of higher order moments

A 4-th order tensor T has an **orthogonal decomposition** if

$$T = \sum_{i=1}^{d} a_i^{\otimes 4}$$

where $\|a_i\| = 1$ and $a_i^T a_j = 0$ for $i \neq j$.

Orthogonal Tensor Decompositon

Motivation

- Latent variable graphical models
- Model parameters are learnt through decomposition of higher order moments

A 4-th order tensor T has an **orthogonal decomposition** if

$$T = \sum_{i=1}^{d} a_i^{\otimes 4}$$

where $\|a_i\| = 1$ and $a_i^T a_j = 0$ for $i \neq j$.

Goal: recover $\{a_i\}$ given T

Orthogonal Tensor Decompositon Challenges

Symmetry

- Inherently susceptible to saddle point issues
- Ask to find d different components a_i, any permutation yields a valid solution
- Creates exponentially many local minima and saddle points

Identify new objective function

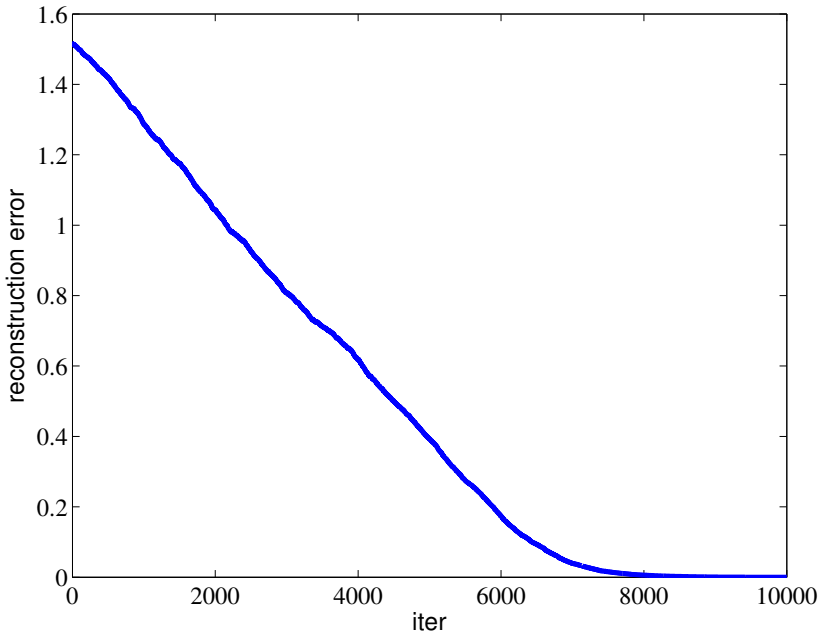
- Satisfies *strict saddle* property
- First online algorithm for orthogonal tensor decomposition with convergence guarantee
- A key step towards making tensor decomposition algorithms scalable

Theorem: Our new objective is strict saddle. Moreover, all its local minima have the form $u_i = \kappa_i a_{\pi(i)}$ for some $\kappa_i = \pm 1$ and permutation $\pi(i)$.

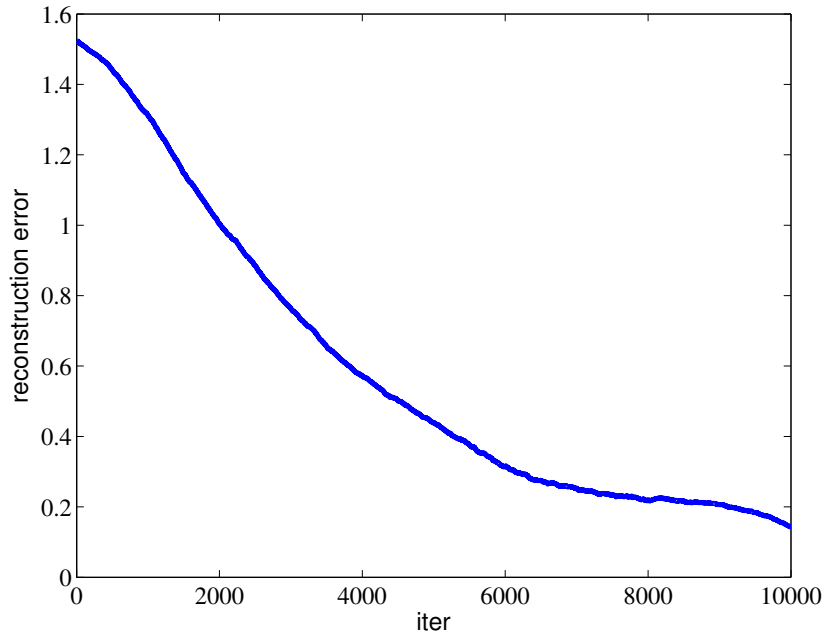
Outline

Experiment

Comparison of different objective functions



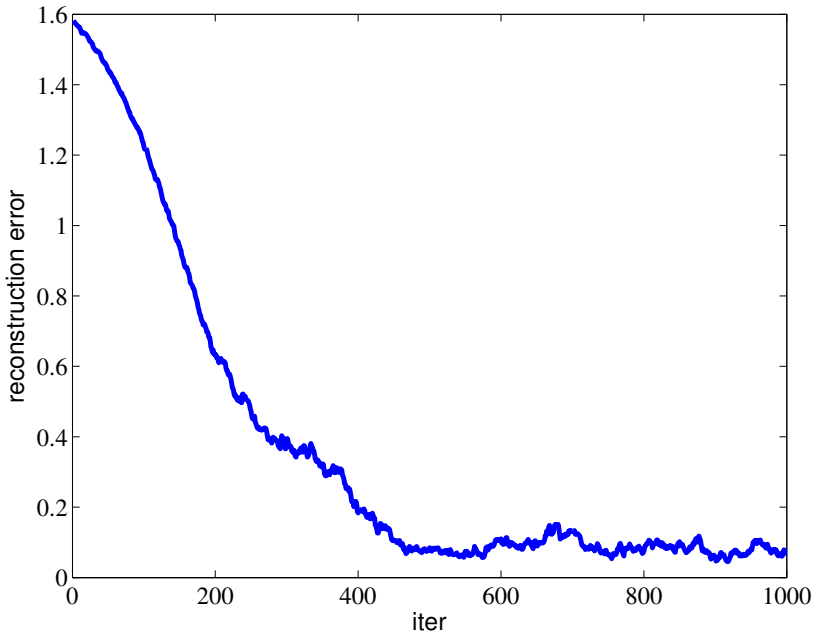
New Objective



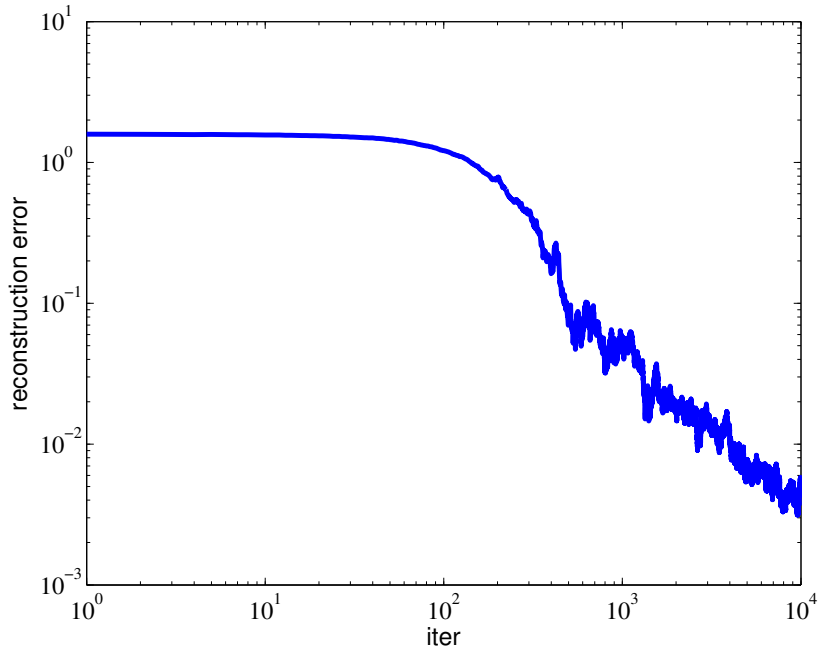
Reconstruction Error Objective

Experiment

ICA setting performance with mini-batch of size 100



Constant Learning Rate η



Learning Rate η/t (in $\log$ scale)

Outline

Conclusion

- Identify the strict saddle property
- Show SGD converges to a local minimum under strict saddle
- New online algorithm for tensor decomposition (non-convex)

Conclusion

- Identify the strict saddle property
- Show SGD converges to a local minimum under strict saddle
- New online algorithm for tensor decomposition (non-convex)

Thank you!

arXiv:1503.02101 [cs.LG]