Mini-Course 1:
SGD Escapes Saddle Points

Yang Yuan

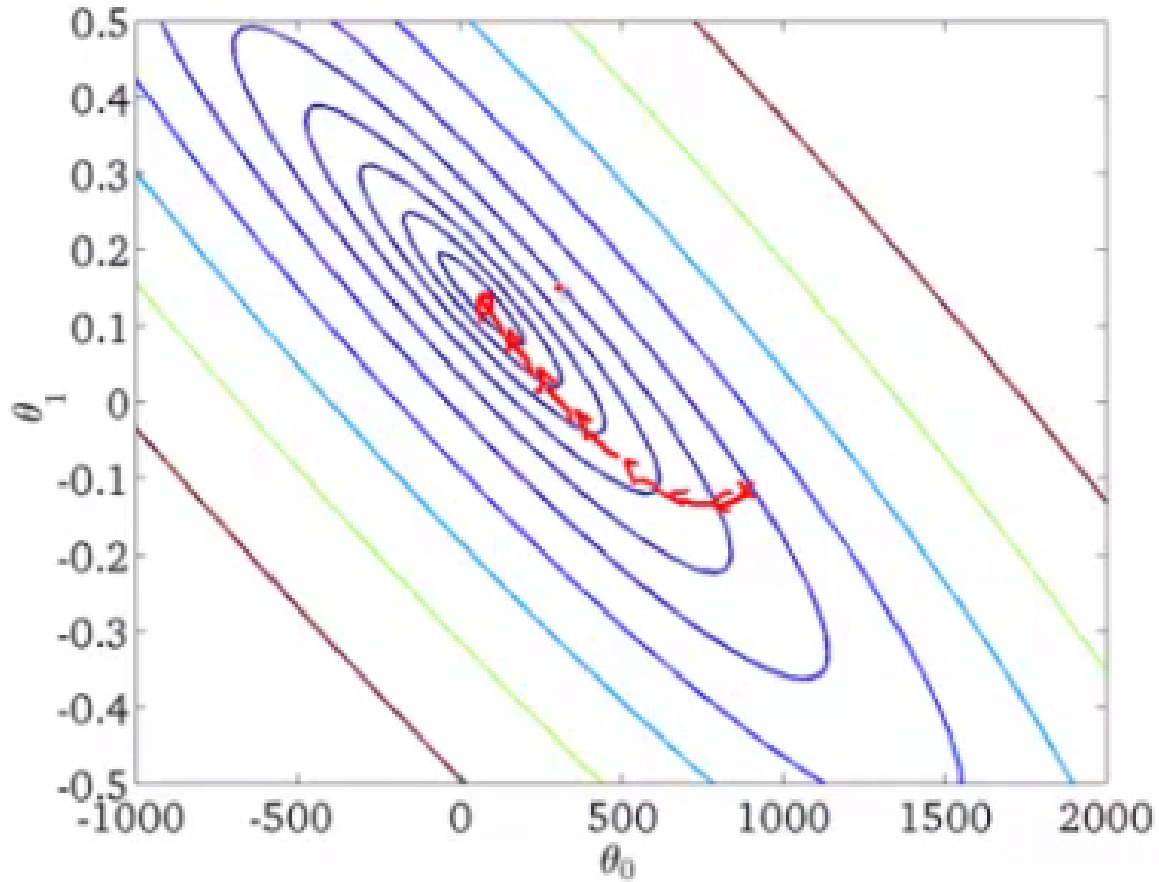
Computer Science Department
Cornell University

Gradient Descent (GD)

- Task: $\min_x f(x)$
- GD does iterative updates $x_{t+1} = x_t - \eta_t \nabla f(x_t)$

Gradient Descent (GD)

- Task: $\min_x f(x)$
- GD does iterative updates $x_{t+1} = x_t - \eta_t \nabla f(x_t)$



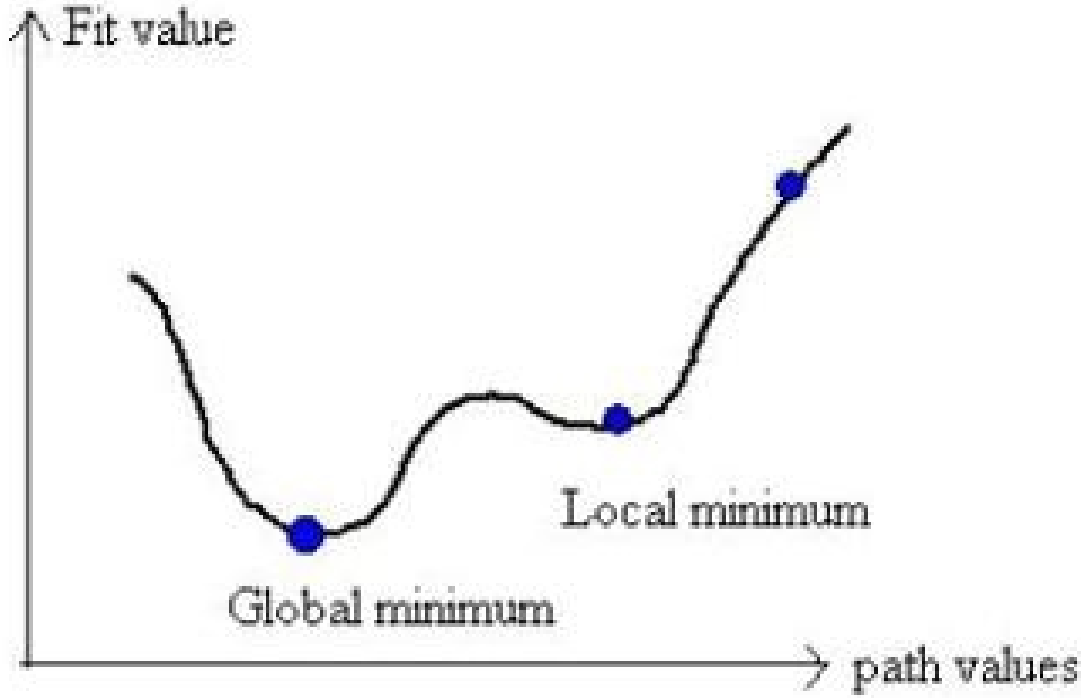
Gradient Descent (GD) has at least two problems

Gradient Descent (GD) has at least two problems

- Computing the full gradient is slow for big data.

Gradient Descent (GD) has at least two problems

- Computing the full gradient is slow for big data.
- Stuck at stationary points.



Stochastic Gradient Descent (SGD)

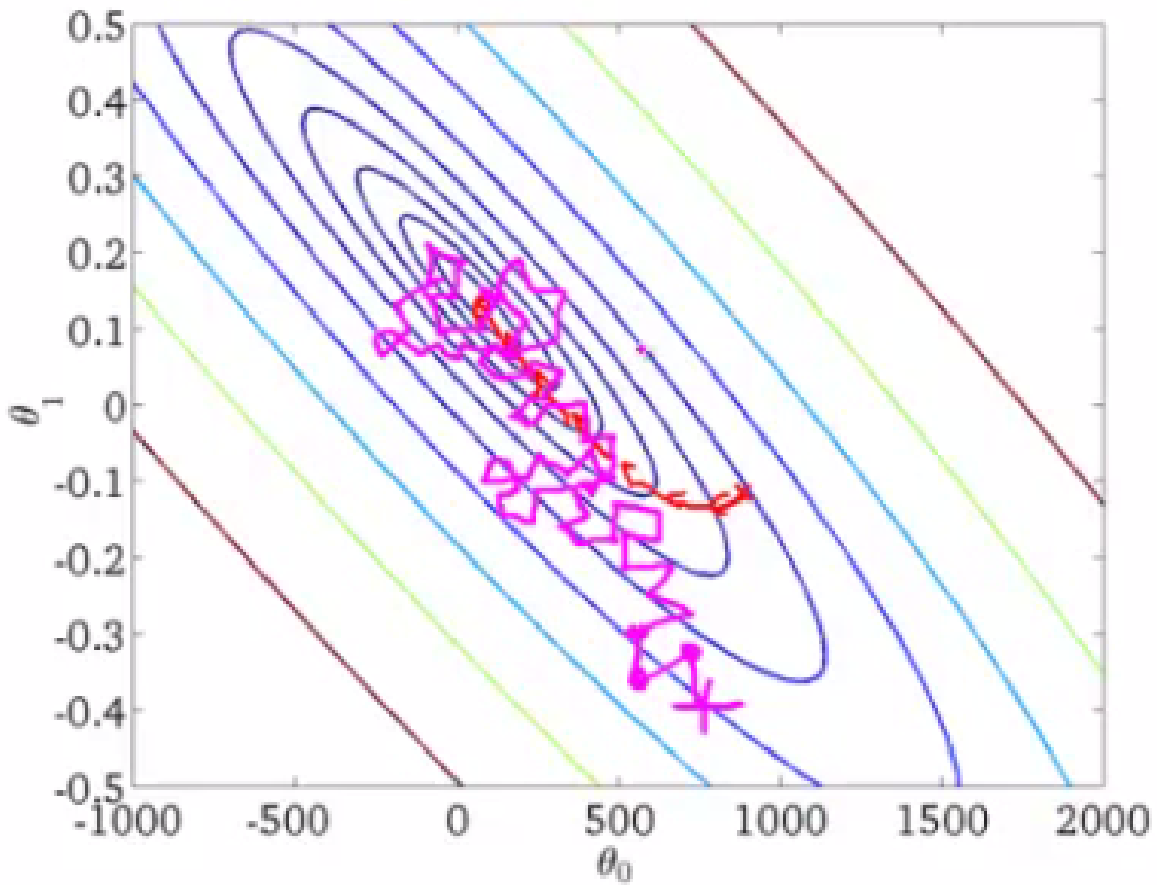
- Very similar to GD, gradient now has some randomness:

$$x_{t+1} = x_t - \eta_t g_t, \text{ where } \mathbb{E}[g_t] = \nabla f(x_t).$$

Stochastic Gradient Descent (SGD)

- Very similar to GD, gradient now has some randomness:

$$x_{t+1} = x_t - \eta_t g_t, \text{ where } \mathbb{E}[g_t] = \nabla f(x_t).$$



Why do we use SGD?

Initially because:

Why do we use SGD?

Initially because:

- Much cheaper to compute using mini-batch

Why do we use SGD?

Initially because:

- Much cheaper to compute using mini-batch
- Can still converge to global minimum in convex case

Why do we use SGD?

Initially because:

- Much cheaper to compute using mini-batch
- Can still converge to global minimum in convex case

But now people realize:

Why do we use SGD?

Initially because:

- Much cheaper to compute using mini-batch
- Can still converge to global minimum in convex case

But now people realize:

- Can escape saddle points! (**Today's topic**)

Why do we use SGD?

Initially because:

- Much cheaper to compute using mini-batch
- Can still converge to global minimum in convex case

But now people realize:

- Can escape saddle points! (**Today's topic**)
- Can escape shallow local minima (**Next time's topic**, some progress.)

Why do we use SGD?

Initially because:

- Much cheaper to compute using mini-batch
- Can still converge to global minimum in convex case

But now people realize:

- Can escape saddle points! (**Today's topic**)
- Can escape shallow local minima (**Next time's topic**, some progress.)
- Can find local minima that generalize well (Not well understood)

Why do we use SGD?

Initially because:

- Much cheaper to compute using mini-batch
- Can still converge to global minimum in convex case

But now people realize:

- Can escape saddle points! (**Today's topic**)
- Can escape shallow local minima (**Next time's topic**, some progress.)
- Can find local minima that generalize well (Not well understood)

Therefore, it's not only faster, but also works better!

About g_t that we use

$$x_{t+1} = x_t - \eta_t g_t, \text{ where } \mathbb{E}[g_t] = \nabla f(x_t).$$

About g_t that we use

$$x_{t+1} = x_t - \eta_t g_t, \text{ where } \mathbb{E}[g_t] = \nabla f(x_t).$$

- In practice, g_t is obtained by sampling a minibatch of size 128 or 256 from the dataset

About g_t that we use

$$x_{t+1} = x_t - \eta_t g_t, \text{ where } \mathbb{E}[g_t] = \nabla f(x_t).$$

- In practice, g_t is obtained by sampling a minibatch of size 128 or 256 from the dataset
- To simplify the analysis, we assume

$$g_t = \nabla f(x_t) + \xi_t$$

where $\xi_t \in N(0, \mathbf{I})$ or $\mathbb{B}_0(r)$

About g_t that we use

$$x_{t+1} = x_t - \eta_t g_t, \text{ where } \mathbb{E}[g_t] = \nabla f(x_t).$$

- In practice, g_t is obtained by sampling a minibatch of size 128 or 256 from the dataset
- To simplify the analysis, we assume

$$g_t = \nabla f(x_t) + \xi_t$$

  where $\xi_t \in N(0, \mathbf{I})$ or $\mathbb{B}_0(r)$
- In general, if ξ_t has non-negligible components on every direction, the analysis works.

Preliminaries

- L-Lipschitz, i.e.,

$$|f(w_1) - f(w_2)| \leq L\|w_1 - w_2\|_2$$

Preliminaries

- L-Lipschitz, i.e.,

$$|f(w_1) - f(w_2)| \leq L\|w_1 - w_2\|_2$$

- ℓ-Smoothness: The gradient is ℓ-Lipschitz, i.e.

$$\|\nabla f(w_1) - \nabla f(w_2)\|_2 \leq \ell\|w_1 - w_2\|_2$$

Preliminaries

- L-Lipschitz, i.e.,

$$|f(w_1) - f(w_2)| \leq L\|w_1 - w_2\|_2$$

- ℓ-Smoothness: The gradient is ℓ-Lipschitz, i.e.

$$\|\nabla f(w_1) - \nabla f(w_2)\|_2 \leq \ell\|w_1 - w_2\|_2$$

- ρ-Hessian smoothness: The hessian matrix is ρ-Lipschitz, i.e.,

$$\|\nabla^2 f(w_1) - \nabla^2 f(w_2)\|_{sp} \leq \rho\|w_1 - w_2\|_2$$

Preliminaries

- L-Lipschitz, i.e.,

$$|f(w_1) - f(w_2)| \leq L\|w_1 - w_2\|_2$$

- ℓ-Smoothness: The gradient is ℓ-Lipschitz, i.e.

$$\|\nabla f(w_1) - \nabla f(w_2)\|_2 \leq \ell\|w_1 - w_2\|_2$$

- ρ-Hessian smoothness: The hessian matrix is ρ-Lipschitz, i.e.,

$$\|\nabla^2 f(w_1) - \nabla^2 f(w_2)\|_{sp} \leq \rho\|w_1 - w_2\|_2$$

  - We need this because we will use the Hessian at the current spot to approximate the neighborhood

Preliminaries

- L-Lipschitz, i.e.,

$$|f(w_1) - f(w_2)| \leq L\|w_1 - w_2\|_2$$

- ℓ-Smoothness: The gradient is ℓ-Lipschitz, i.e.

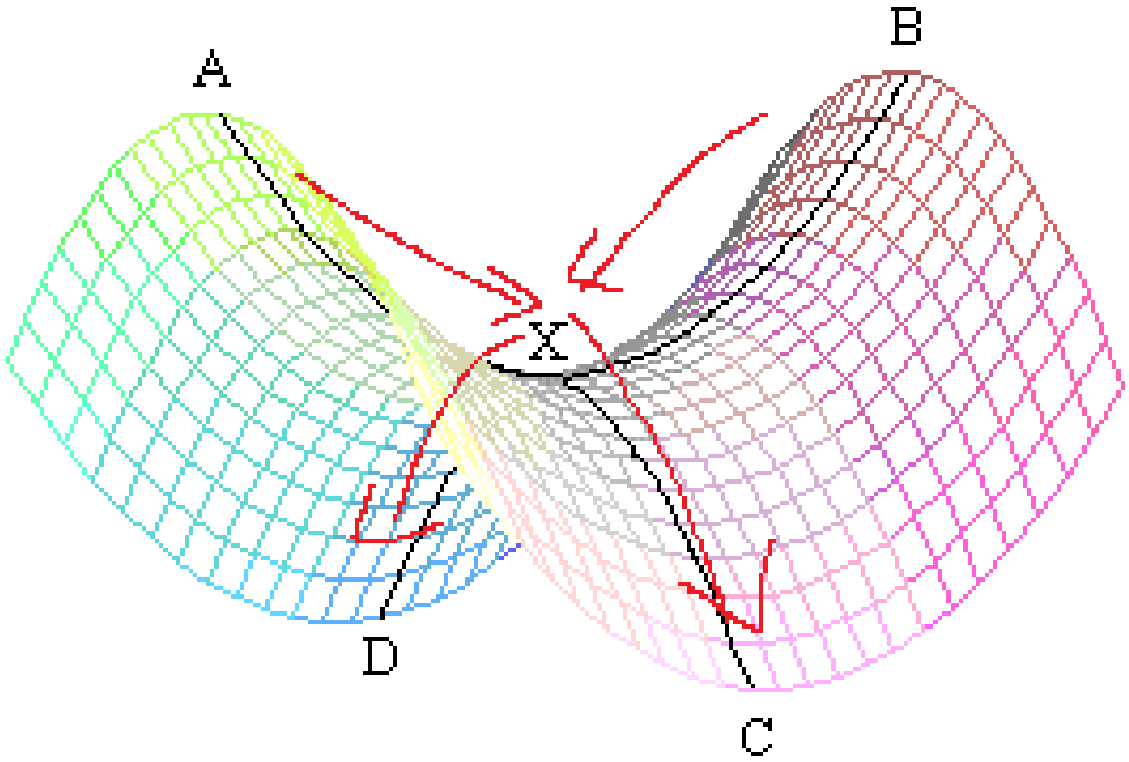
$$\|\nabla f(w_1) - \nabla f(w_2)\|_2 \leq \ell\|w_1 - w_2\|_2$$

- ρ-Hessian smoothness: The hessian matrix is ρ-Lipschitz, i.e.,

$$\|\nabla^2 f(w_1) - \nabla^2 f(w_2)\|_{sp} \leq \rho\|w_1 - w_2\|_2$$

  - We need this because we will use the Hessian at the current spot to approximate the neighborhood
  - Then bound the approximation.

Saddle points, and negative eigenvalue



Stationary points: saddle points, local minima, local maxima

For stationary points $\nabla f(w) = 0$,

Stationary points: saddle points, local minima, local maxima

For stationary points $\nabla f(w) = 0$,

- If $\nabla^2 f(w) \succ 0$, it's a local minimum.

Stationary points: saddle points, local minima, local maxima

For stationary points $\nabla f(w) = 0$,

- If $\nabla^2 f(w) \succ 0$, it's a local minimum.
- If $\nabla^2 f(w) \prec 0$, it's a local maximum.

Stationary points: saddle points, local minima, local maxima

For stationary points $\nabla f(w) = 0$,

- If $\nabla^2 f(w) \succ 0$, it's a local minimum.
- If $\nabla^2 f(w) \prec 0$, it's a local maximum.
- If $\nabla^2 f(w)$ has both $+/-$ eigenvalues, it's a saddle point.

Stationary points: saddle points, local minima, local maxima

For stationary points $\nabla f(w) = 0$,

- If $\nabla^2 f(w) \succ 0$, it's a local minimum.
- If $\nabla^2 f(w) \prec 0$, it's a local maximum.
- If $\nabla^2 f(w)$ has both $+/-$ eigenvalues, it's a saddle point.
- **Degenerate case:** $\nabla^2 f(w)$ has eigenvalues equal to 0. It could be either local minimum(maximum)/saddle point.

Stationary points: saddle points, local minima, local maxima

For stationary points $\nabla f(w) = 0$,

- If $\nabla^2 f(w) \succ 0$, it's a local minimum.
- If $\nabla^2 f(w) \prec 0$, it's a local maximum.
- If $\nabla^2 f(w)$ has both $+/-$ eigenvalues, it's a saddle point.
- **Degenerate case:** $\nabla^2 f(w)$ has eigenvalues equal to 0. It could be either local minimum(maximum)/saddle point.
  - f is "flat" on some directions

Stationary points: saddle points, local minima, local maxima

For stationary points $\nabla f(w) = 0$,

- If $\nabla^2 f(w) \succ 0$, it's a local minimum.
- If $\nabla^2 f(w) \prec 0$, it's a local maximum.
- If $\nabla^2 f(w)$ has both $+/-$ eigenvalues, it's a saddle point.
- **Degenerate case:** $\nabla^2 f(w)$ has eigenvalues equal to 0. It could be either local minimum(maximum)/saddle point.
  - f is "flat" on some directions
  - SGD is like random walk

Stationary points: saddle points, local minima, local maxima

For stationary points $\nabla f(w) = 0$,

- If $\nabla^2 f(w) \succ 0$, it's a local minimum.
- If $\nabla^2 f(w) \prec 0$, it's a local maximum.
- If $\nabla^2 f(w)$ has both $+/-$ eigenvalues, it's a saddle point.
- **Degenerate case:** $\nabla^2 f(w)$ has eigenvalues equal to 0. It could be either local minimum(maximum)/saddle point.
  - f is "flat" on some directions
  - SGD is like random walk
  - We only consider non-degenerate case!

Strict saddle property

$f(w)$ is $(\alpha, \gamma, \epsilon, \zeta)$-strict saddle, if for any w,

Strict saddle property

$f(w)$ is $(\alpha, \gamma, \epsilon, \zeta)$-strict saddle, if for any w,

- $\|\nabla f(w)\|_2 \geq \epsilon$

Which means:

- Gradient is large

Strict saddle property

$f(w)$ is $(\alpha, \gamma, \epsilon, \zeta)$-strict saddle, if for any w,

- $\|\nabla f(w)\|_2 \geq \epsilon$
- or, $\lambda_{\min} \nabla^2 f(w) \leq -\gamma < 0$

Which means:

- Gradient is large
- or (stationary point), we have a negative eigenvalue direction to escape

Strict saddle property

$f(w)$ is $(\alpha, \gamma, \epsilon, \zeta)$-strict saddle, if for any w,

- $\|\nabla f(w)\|_2 \geq \epsilon$
- or, $\lambda_{\min} \nabla^2 f(w) \leq -\gamma < 0$
- or, there exists w^* such that $\|w - w^*\|_2 \leq \zeta$, and the region centered w^* with radius 2ζ is α-strongly convex.

Which means:

- Gradient is large
- or (stationary point), we have a negative eigenvalue direction to escape
- or (stationary point, no negative eigenvalues), we are pretty close to a local minimum.

Strict saddle functions are everywhere

- Orthogonal tensor decomposition [Ge et al 2015]
- Deep linear (residual) networks [Kawaguchi 2016], [Hardt and Ma 2016]
- Matrix completion [Ge et al 2016]
- Generalized phase retrieval problem [Sun et al 2016]
- Low rank matrix recovery [Bhojanapalli et al 2016]

Strict saddle functions are everywhere

- Orthogonal tensor decomposition [Ge et al 2015]
- Deep linear (residual) networks [Kawaguchi 2016], [Hardt and Ma 2016]
- Matrix completion [Ge et al 2016]
- Generalized phase retrieval problem [Sun et al 2016]
- Low rank matrix recovery [Bhojanapalli et al 2016]

Moreover, in these problems, all local minima are equally good!

Strict saddle functions are everywhere

- Orthogonal tensor decomposition [Ge et al 2015]
- Deep linear (residual) networks [Kawaguchi 2016], [Hardt and Ma 2016]
- Matrix completion [Ge et al 2016]
- Generalized phase retrieval problem [Sun et al 2016]
- Low rank matrix recovery [Bhojanapalli et al 2016]

Moreover, in these problems, all local minima are equally good! That means,

Strict saddle functions are everywhere

- Orthogonal tensor decomposition [Ge et al 2015]
- Deep linear (residual) networks [Kawaguchi 2016], [Hardt and Ma 2016]
- Matrix completion [Ge et al 2016]
- Generalized phase retrieval problem [Sun et al 2016]
- Low rank matrix recovery [Bhojanapalli et al 2016]

Moreover, in these problems, all local minima are equally good! That means,

- SGD escapes all saddle points

Strict saddle functions are everywhere

- Orthogonal tensor decomposition [Ge et al 2015]
- Deep linear (residual) networks [Kawaguchi 2016], [Hardt and Ma 2016]
- Matrix completion [Ge et al 2016]
- Generalized phase retrieval problem [Sun et al 2016]
- Low rank matrix recovery [Bhojanapalli et al 2016]

Moreover, in these problems, all local minima are equally good! That means,

- SGD escapes all saddle points
- So, SGD arrives one local minimum $\rightarrow$ global minimum!

Strict saddle functions are everywhere

- Orthogonal tensor decomposition [Ge et al 2015]
- Deep linear (residual) networks [Kawaguchi 2016], [Hardt and Ma 2016]
- Matrix completion [Ge et al 2016]
- Generalized phase retrieval problem [Sun et al 2016]
- Low rank matrix recovery [Bhojanapalli et al 2016]

Moreover, in these problems, all local minima are equally good! That means,

- SGD escapes all saddle points
- So, SGD arrives one local minimum $\rightarrow$ global minimum!
- One popular way to prove SGD solves the problem.

Main Results

- [Ge et al 2015] says, whp, SGD will escape all saddle points, and converge to a local minimum. The convergence time has polynomial dependency in dimension d.

Main Results

- [Ge et al 2015] says, whp, SGD will escape all saddle points, and converge to a local minimum. The convergence time has polynomial dependency in dimension d.
- [Jin et al 2017] says, whp, PGD (a variant of SGD) will escape all saddle points, and converge to a local minimum much faster. The dependence in d is **logarithmic**.

Main Results

- [Ge et al 2015] says, whp, SGD will escape all saddle points, and converge to a local minimum. The convergence time has polynomial dependency in dimension d.
- [Jin et al 2017] says, whp, PGD (a variant of SGD) will escape all saddle points, and converge to a local minimum much faster. The dependence in d is **logarithmic**.
- Same proof framework. We'll mainly look at the new result.

Description of PGD

Do the following iteratively:

Description of PGD

Do the following iteratively:

- Do a gradient descent step:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

Description of PGD

Do the following iteratively:

- If $\|\nabla f(x_t)\| \leq g_{\text{thres}}$, and last perturbed time is $> t_{\text{thres}}$ steps before, do random perturbation (ball)

- Do a gradient descent step:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

Description of PGD

Do the following iteratively:

- If $\|\nabla f(x_t)\| \leq g_{\text{thres}}$, and last perturbed time is $> t_{\text{thres}}$ steps before, do random perturbation (ball)
- If perturbation happened t_{thres} steps ago, but f is decreased for less than f_{thres}, return the value before last perturbation
- Do a gradient descent step:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

Description of PGD

Do the following iteratively:

- If $\|\nabla f(x_t)\| \leq g_{\text{thres}}$, and last perturbed time is $> t_{\text{thres}}$ steps before, do random perturbation (ball)
- If perturbation happened t_{thres} steps ago, but f is decreased for less than f_{thres}, return the value before last perturbation
- Do a gradient descent step:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

A few Remarks:

Description of PGD

Do the following iteratively:

- If $\|\nabla f(x_t)\| \leq g_{\text{thres}}$, and last perturbed time is $> t_{\text{thres}}$ steps before, do random perturbation (ball)
- If perturbation happened t_{thres} steps ago, but f is decreased for less than f_{thres}, return the value before last perturbation
- Do a gradient descent step:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

A few Remarks:

- Unfortunately.. Not a fast algorithm because of GD!

Description of PGD

Do the following iteratively:

- If $\|\nabla f(x_t)\| \leq g_{\text{thres}}$, and last perturbed time is $> t_{\text{thres}}$ steps before, do random perturbation (ball)
- If perturbation happened t_{thres} steps ago, but f is decreased for less than f_{thres}, return the value before last perturbation
- Do a gradient descent step:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

A few Remarks:

- Unfortunately.. Not a fast algorithm because of GD!
- $\eta = \frac{c}{\ell}$. $g_{\text{thres}}, t_{\text{thres}}, f_{\text{thres}}$ depends on a constant c, as well as other parameters.

Main theorem in [Jin et al 2017]

Theorem (Main Theorem)

Assume function f is ℓ-smooth and ρ-Hessian Lipschitz, $(\alpha, \gamma, \epsilon, \zeta)$-strict saddle. There exists an absolute constant $c_{\max}$ such that, for any $\delta > 0$, $\Delta_f \geq f(x_0) - f^*$, and constant $c \leq c_{\max}$, $\tilde{\epsilon} = \min\{\epsilon, \frac{\gamma^2}{\rho}\}$, PGD$(c)$ will output a point ζ-close to a local minimum, with probability $1 - \delta$, and terminate in the following number of iterations:

$$O\left(\frac{\ell(f(x_0) - f^*)}{\tilde{\epsilon}^2} \log^4\left(\frac{d\ell\Delta_f}{\tilde{\epsilon}^2\delta}\right)\right)$$

Main theorem in [Jin et al 2017]

Theorem (Main Theorem)

Assume function f is ℓ-smooth and ρ-Hessian Lipschitz, $(\alpha, \gamma, \epsilon, \zeta)$-strict saddle. There exists an absolute constant $c_{\max}$ such that, for any $\delta > 0$, $\Delta_f \geq f(x_0) - f^*$, and constant $c \leq c_{\max}$, $\tilde{\epsilon} = \min\{\epsilon, \frac{\gamma^2}{\rho}\}$, PGD$(c)$ will output a point ζ-close to a local minimum, with probability $1 - \delta$, and terminate in the following number of iterations:

$$O\left(\frac{\ell(f(x_0) - f^*)}{\tilde{\epsilon}^2} \log^4\left(\frac{d\ell\Delta_f}{\tilde{\epsilon}^2\delta}\right)\right)$$

- If could show SGD has similar property, would be great!

Main theorem in [Jin et al 2017]

Theorem (Main Theorem)

Assume function f is ℓ-smooth and ρ-Hessian Lipschitz, $(\alpha, \gamma, \epsilon, \zeta)$-strict saddle. There exists an absolute constant $c_{\max}$ such that, for any $\delta > 0$, $\Delta_f \geq f(x_0) - f^*$, and constant $c \leq c_{\max}$, $\tilde{\epsilon} = \min\{\epsilon, \frac{\gamma^2}{\rho}\}$, PGD$(c)$ will output a point ζ-close to a local minimum, with probability $1 - \delta$, and terminate in the following number of iterations:

$$O\left(\frac{\ell(f(x_0) - f^*)}{\tilde{\epsilon}^2} \log^4\left(\frac{d\ell\Delta_f}{\tilde{\epsilon}^2\delta}\right)\right)$$

- If could show SGD has similar property, would be great!
- The convergence rate is almost optimal.

More general version: why it's fast

Theorem (A more general version)

Assume function f is ℓ-smooth and ρ-Hessian Lipschitz. There exists an absolute constant $c_{\max}$ such that, for any $\delta > 0$, $\Delta_f \geq f(x_0) - f^*$, and constant $c \leq c_{\max}$, $\tilde{\epsilon} \leq \frac{\ell^2}{\rho}$, PGD$(c)$ will output a point ζ-close to an $\tilde{\epsilon}$-second-order stationary point, with probability $1 - \delta$, and terminate in the following number of iterations:

$$O\left(\frac{\ell(f(x_0) - f^*)}{\tilde{\epsilon}^2} \log^4\left(\frac{d\ell\Delta_f}{\tilde{\epsilon}^2\delta}\right)\right)$$

Essentially saying the **same** thing. If f is not strict saddle, only ϵ-second-order stationary point (instead of local minimum) is guaranteed.

ϵ-stationary points

- ϵ-first-order stationary point:

$$\|\nabla f(x)\| \leq \epsilon$$

ϵ-stationary points

- ϵ-first-order stationary point:

$$\|\nabla f(x)\| \leq \epsilon$$

- ϵ-second-order stationary point:

$$\|\nabla f(x)\| \leq \epsilon, \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\rho\epsilon}$$

ϵ-stationary points

- ϵ-first-order stationary point:

$$\|\nabla f(x)\| \leq \epsilon$$

- ϵ-second-order stationary point:

$$\|\nabla f(x)\| \leq \epsilon, \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\rho\epsilon}$$

  - If ℓ-smooth, $\lambda_{\min}(\nabla^2 f(x)) \geq -\ell$.

ϵ-stationary points

- ϵ-first-order stationary point:

$$\|\nabla f(x)\| \leq \epsilon$$

- ϵ-second-order stationary point:

$$\|\nabla f(x)\| \leq \epsilon, \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\rho\epsilon}$$

  - If ℓ-smooth, $\lambda_{\min}(\nabla^2 f(x)) \geq -\ell$.
  - For any $\epsilon > \frac{\ell^2}{\rho}$, an ϵ-first-order stationary point in a ℓ-smooth function is a $\frac{\ell^2}{\rho}$-second-order stationary point

ϵ-stationary points

- ϵ-first-order stationary point:

$$\|\nabla f(x)\| \leq \epsilon$$

- ϵ-second-order stationary point:

$$\|\nabla f(x)\| \leq \epsilon, \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\rho\epsilon}$$

  - If ℓ-smooth, $\lambda_{\min}(\nabla^2 f(x)) \geq -\ell$.
  - For any $\epsilon > \frac{\ell^2}{\rho}$, an ϵ-first-order stationary point in a ℓ-smooth function is a $\frac{\ell^2}{\rho}$-second-order stationary point
  - If $(\alpha, \gamma, \epsilon, \zeta)$-strict saddle, and $\epsilon < \frac{\gamma^2}{\rho}$, then any ϵ-second-order stationary point is a local minimum.

[Nesterov, 1998]

Theorem

Assume that f is ℓ-smooth. Then for any $\tilde{\epsilon} > 0$, if we run GD with step size $\eta = \frac{1}{\ell}$ and termination condition $\|\nabla f(x)\| \leq \tilde{\epsilon}$, the output will be $\tilde{\epsilon}$-first-order stationary point, and the algorithm terminates in the following number of iterations:

$$\frac{\ell(f(x_0) - f^*)}{\tilde{\epsilon}^2}$$

[Jin et al 2017]: PGD converges to $\tilde{\epsilon}$-second-order stationary point in $O\left(\frac{\ell(f(x_0)-f^*)}{\tilde{\epsilon}^2} \log^4\left(\frac{d\ell\Delta_f}{\tilde{\epsilon}^2\delta}\right)\right)$ steps.

- Matched up to log factors!

Why $-\sqrt{\rho\epsilon}$?

- If we use third order approximation for x [Nesterov and Polyak, 2006]

$$\min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{\rho}{6} \|y - x\|^2 \right\}$$

denote the answer as T_x.

Why $-\sqrt{\rho\epsilon}$?

- If we use third order approximation for x [Nesterov and Polyak, 2006]

$$\min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{\rho}{6} \|y - x\|^2 \right\}$$

  denote the answer as T_x.
- Denote distance $r = \|x - T_x\|$

Why $-\sqrt{\rho\epsilon}$?

- If we use third order approximation for x [Nesterov and Polyak, 2006]

$$\min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{\rho}{6} \|y - x\|^2 \right\}$$

  denote the answer as T_x.
- Denote distance $r = \|x - T_x\|$
- $\|\nabla f(T_x)\| \leq \rho r^2$, $\nabla^2 f(T_x) \succ \frac{-3}{2} \rho r \mathbf{I}$

Why $-\sqrt{\rho\epsilon}$?

- If we use third order approximation for x [Nesterov and Polyak, 2006]

$$\min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{\rho}{6} \|y - x\|^2 \right\}$$

  denote the answer as T_x.
- Denote distance $r = \|x - T_x\|$
- $\|\nabla f(T_x)\| \leq \rho r^2$, $\nabla^2 f(T_x) \succ \frac{-3}{2} \rho r \mathbf{I}$
- To get a lower bound for r:

$$\max \left\{ \sqrt{\frac{\|\nabla f(T_x)\|}{\rho}}, -\frac{2}{3\rho} \lambda_{\min} \nabla^2 f(T_x) \right\}$$

Why $-\sqrt{\rho\epsilon}$?

- If we use third order approximation for x [Nesterov and Polyak, 2006]

$$\min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{\rho}{6} \|y - x\|^2 \right\}$$

  denote the answer as T_x.
- Denote distance $r = \|x - T_x\|$
- $\|\nabla f(T_x)\| \leq \rho r^2$, $\nabla^2 f(T_x) \succ \frac{-3}{2} \rho r \mathbf{I}$
- To get a lower bound for r:

$$\max \left\{ \sqrt{\frac{\|\nabla f(T_x)\|}{\rho}}, -\frac{2}{3\rho} \lambda_{\min} \nabla^2 f(T_x) \right\}$$

- When are they equal $\rightarrow -\sqrt{\rho\epsilon}$

Related results

1. “Gradient Descent Converges to Minimizers” By Lee, Simchowitz, Jordan and Recht. 15’

Related results

1. "Gradient Descent Converges to Minimizers" By Lee, Simchowitz, Jordan and Recht. 15'
   - with random initialization, GD almost surely never touches any saddle points, and always converges to local minima.

Related results

1. "Gradient Descent Converges to Minimizers" By Lee, Simchowitz, Jordan and Recht. 15'
   - with random initialization, GD almost surely never touches any saddle points, and always converges to local minima.
2. "The power of normalization: faster evasion of saddle points", Kfir Levy. 16'

Related results

1. "Gradient Descent Converges to Minimizers" By Lee, Simchowitz, Jordan and Recht. 15'
   - with random initialization, GD almost surely never touches any saddle points, and always converges to local minima.
2. "The power of normalization: faster evasion of saddle points", Kfir Levy. 16'
   - Normalized gradient can escape saddle points in $O(d^3 poly(1/\epsilon))$, slower than [Jin et al 2017], faster than [Ge et al 2015], but still polynomial in d.

Main theorem in [Jin et al 2017]

Theorem (Main Theorem)

Assume function f is ℓ-smooth and ρ-Hessian Lipschitz, $(\alpha, \gamma, \epsilon, \zeta)$-strict saddle. There exists an absolute constant $c_{\max}$ such that, for any $\delta > 0$, $\Delta_f \geq f(x_0) - f^*$, and constant $c \leq c_{\max}$, $\tilde{\epsilon} = \min\{\epsilon, \frac{\gamma^2}{\rho}\}$, PGD$(c)$ will output a point ζ-close to a local minimum, with probability $1 - \delta$, and terminate in the following number of iterations:

$$O\left(\frac{\ell(f(x_0) - f^*)}{\tilde{\epsilon}^2} \log^4\left(\frac{d\ell\Delta_f}{\tilde{\epsilon}^2\delta}\right)\right)$$

Proof framework: Progress, Escape and Trap

- **Progress**: when $\|\nabla f(x)\| > g_{\text{thres}}$, $f(x)$ is decreased by at least $f_{\text{thres}}/t_{\text{thres}}$.

Proof framework: Progress, Escape and Trap

- **Progress**: when $\|\nabla f(x)\| > g_{\text{thres}}$, $f(x)$ is decreased by at least $f_{\text{thres}}/t_{\text{thres}}$.
- **Escape**: when $\|\nabla f(x)\| \leq g_{\text{thres}}$, and $\lambda_{\min}\nabla^2 f(x) \leq -\gamma$, whp function value is decreased by f_{thres} after perturbation$+t_{\text{thres}}$ steps.

Proof framework: Progress, Escape and Trap

- **Progress**: when $\|\nabla f(x)\| > g_{\text{thres}}$, $f(x)$ is decreased by at least $f_{\text{thres}}/t_{\text{thres}}$.
- **Escape**: when $\|\nabla f(x)\| \leq g_{\text{thres}}$, and $\lambda_{\min}\nabla^2 f(x) \leq -\gamma$, whp function value is decreased by f_{thres} after perturbation$+t_{\text{thres}}$ steps.
  - $f_{\text{thres}}/t_{\text{thres}}$ on average each step.

Proof framework: Progress, Escape and Trap

- **Progress**: when $\|\nabla f(x)\| > g_{\text{thres}}$, $f(x)$ is decreased by at least $f_{\text{thres}}/t_{\text{thres}}$.
- **Escape**: when $\|\nabla f(x)\| \leq g_{\text{thres}}$, and $\lambda_{\min}\nabla^2 f(x) \leq -\gamma$, whp function value is decreased by f_{thres} after perturbation$+t_{\text{thres}}$ steps.
  - $f_{\text{thres}}/t_{\text{thres}}$ on average each step.
- **Trap**:

Proof framework: Progress, Escape and Trap

- **Progress**: when $\|\nabla f(x)\| > g_{\text{thres}}$, $f(x)$ is decreased by at least $f_{\text{thres}}/t_{\text{thres}}$.
- **Escape**: when $\|\nabla f(x)\| \leq g_{\text{thres}}$, and $\lambda_{\min}\nabla^2 f(x) \leq -\gamma$, whp function value is decreased by f_{thres} after perturbation$+t_{\text{thres}}$ steps.
  - $f_{\text{thres}}/t_{\text{thres}}$ on average each step.
- **Trap**:
  - The algorithm can't do progress and escape forever, because it's bounded!

Proof framework: Progress, Escape and Trap

- **Progress**: when $\|\nabla f(x)\| > g_{\text{thres}}$, $f(x)$ is decreased by at least $f_{\text{thres}}/t_{\text{thres}}$.
- **Escape**: when $\|\nabla f(x)\| \leq g_{\text{thres}}$, and $\lambda_{\min}\nabla^2 f(x) \leq -\gamma$, whp function value is decreased by f_{thres} after perturbation$+t_{\text{thres}}$ steps.
  - $f_{\text{thres}}/t_{\text{thres}}$ on average each step.
- **Trap**:
  - The algorithm can't do progress and escape forever, because it's bounded!
  - When it stops: perturbation happened t_{thres} steps ago, but f is decreased for less than f_{thres}

Proof framework: Progress, Escape and Trap

- **Progress**: when $\|\nabla f(x)\| > g_{\text{thres}}$, $f(x)$ is decreased by at least $f_{\text{thres}}/t_{\text{thres}}$.
- **Escape**: when $\|\nabla f(x)\| \leq g_{\text{thres}}$, and $\lambda_{\min}\nabla^2 f(x) \leq -\gamma$, whp function value is decreased by f_{thres} after perturbation$+t_{\text{thres}}$ steps.
  - $f_{\text{thres}}/t_{\text{thres}}$ on average each step.
- **Trap**:
  - The algorithm can't do progress and escape forever, because it's bounded!
  - When it stops: perturbation happened t_{thres} steps ago, but f is decreased for less than f_{thres}
  - That means, $\|\nabla f(x)\| \leq g_{\text{thres}}$ before perturbation, and whp there is no eigenvalue $\leq -\gamma$.

Proof framework: Progress, Escape and Trap

- **Progress**: when $\|\nabla f(x)\| > g_{\text{thres}}$, $f(x)$ is decreased by at least $f_{\text{thres}}/t_{\text{thres}}$.
- **Escape**: when $\|\nabla f(x)\| \leq g_{\text{thres}}$, and $\lambda_{\min}\nabla^2 f(x) \leq -\gamma$, whp function value is decreased by f_{thres} after perturbation$+t_{\text{thres}}$ steps.
  - $f_{\text{thres}}/t_{\text{thres}}$ on average each step.
- **Trap**:
  - The algorithm can't do progress and escape forever, because it's bounded!
  - When it stops: perturbation happened t_{thres} steps ago, but f is decreased for less than f_{thres}
  - That means, $\|\nabla f(x)\| \leq g_{\text{thres}}$ before perturbation, and whp there is no eigenvalue $\leq -\gamma$.
  - So it's a local minimum!

Progress

Lemma

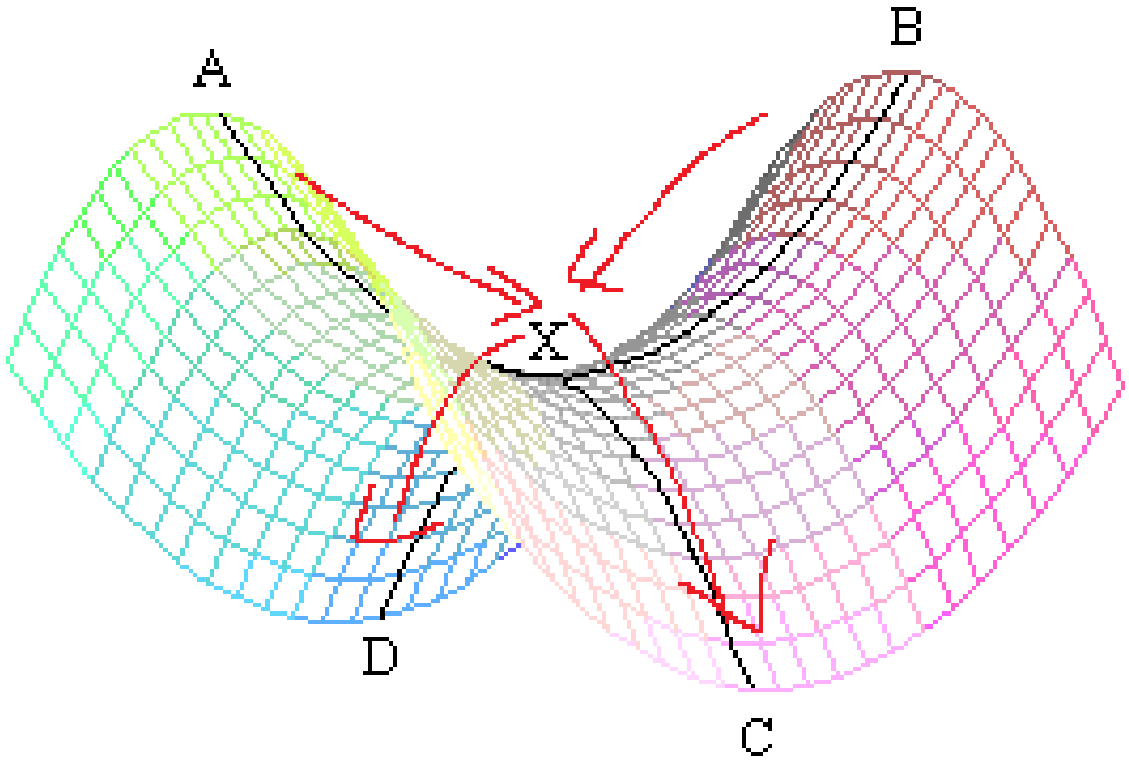
If f is ℓ-smooth, then for GD with step size $\eta < \frac{1}{\ell}$, we have:

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2$$

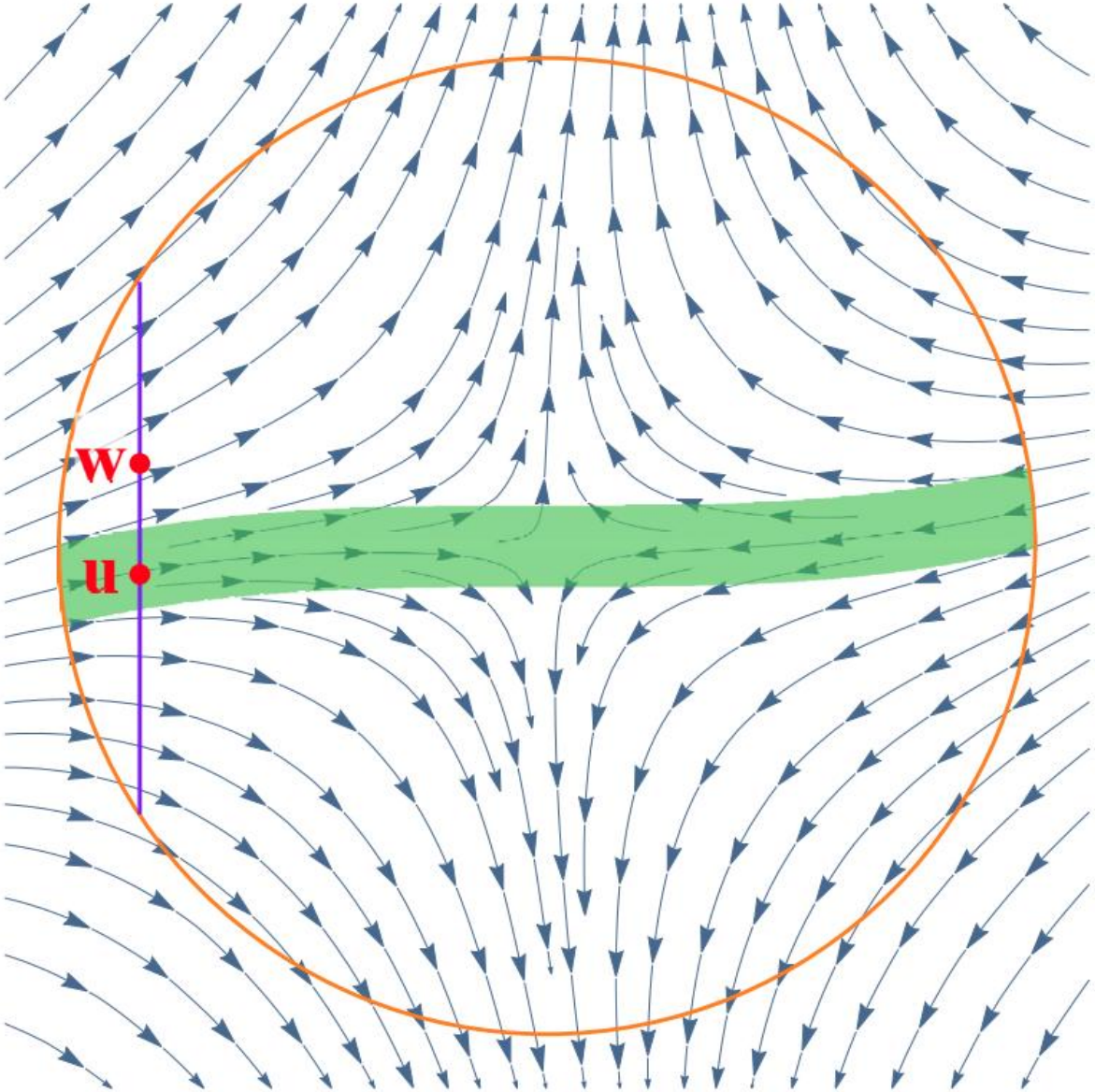
Proof.

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{\ell}{2}\|x_{t+1} - x_t\|^2 \\ &= f(x_t) - \eta\|\nabla f(x_t)\|^2 + \frac{\eta^2 \ell}{2}\|\nabla f(x_t)\|^2 \\ &\leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2 \end{aligned}$$

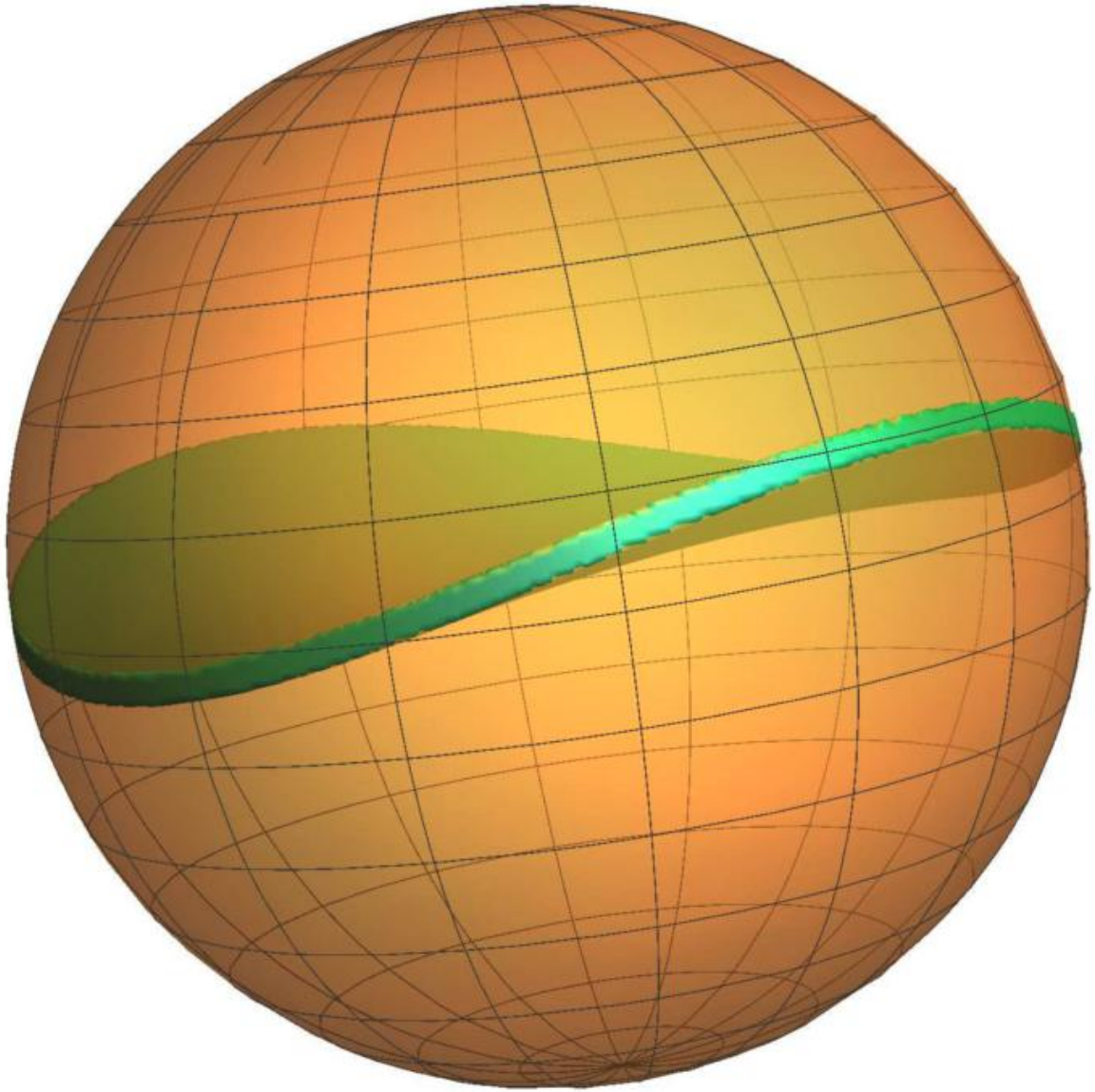
Escape: main idea



Escape: main idea



Escape: thin pancake



Main Lemma: measure the width

Lemma

Suppose we start with point $\tilde{x}$ satisfying following conditions:

$$\|\nabla f(\tilde{x})\| \leq g_{\text{thres}}, \qquad \lambda_{\min}(\nabla^2 f(\tilde{x})) \leq -\gamma$$

Let e_1 the minimum eigenvector. Consider two gradient descent sequences $\{u_t\}, \{w_t\}$, with initial points u_0, w_0 satisfying :

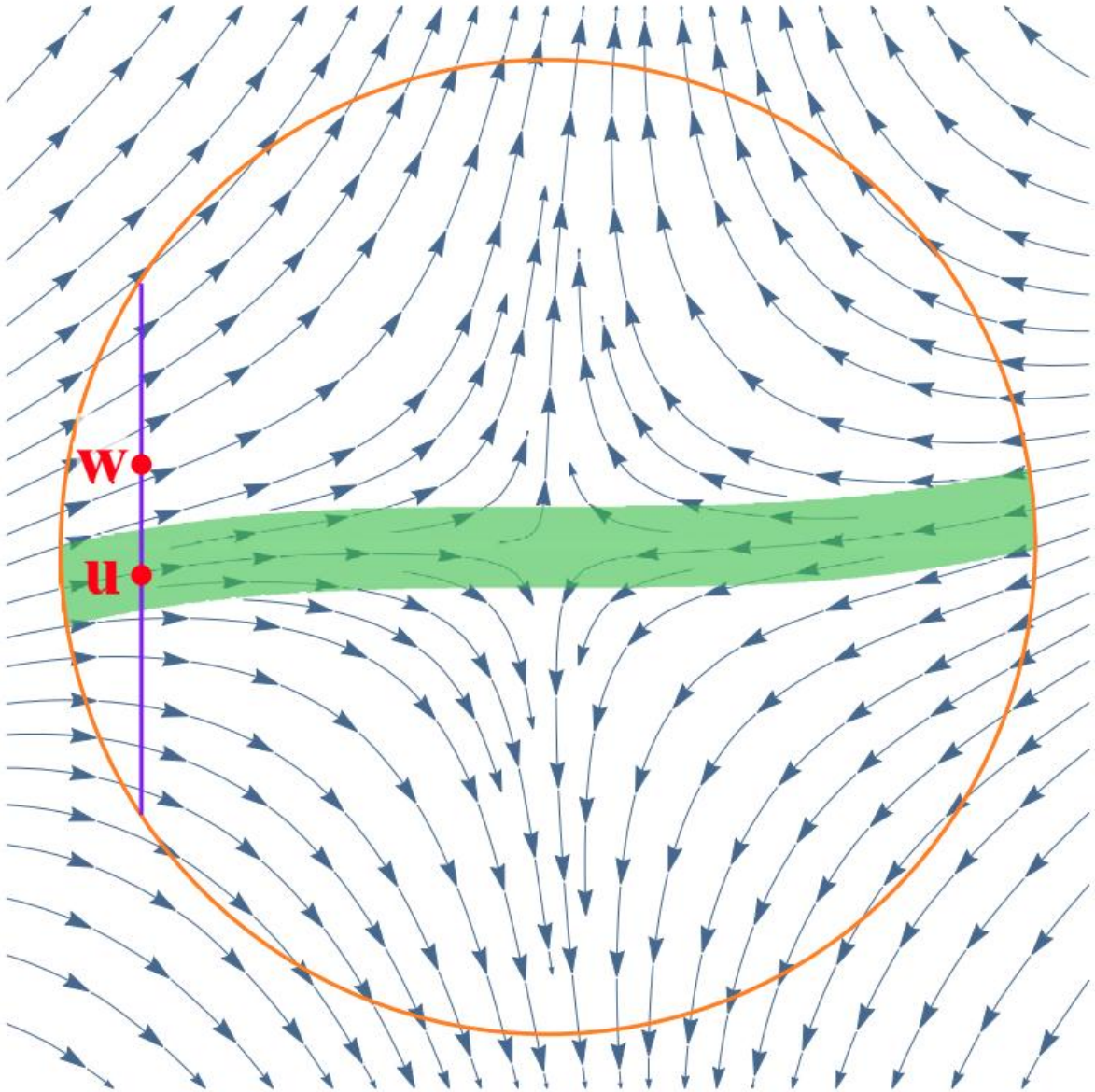
$$\|u_0 - \tilde{x}\| \leq r, w_0 = u_0 + \mu r e_1, \mu \in [\delta/(2\sqrt{d}), 1]$$

Then, for any stepsize $\eta \leq c_{\max}/\ell$, and any $T \geq t_{\text{thres}}$, we have

$$\min\{f(u_T) - f(u_0), f(w_T) - f(w_0)\} \leq -2.5 f_{\text{thres}}$$

- As long as $u_0 - w_0$ are on e_1, and $\|u_0 - w_0\| \geq \frac{\delta r}{2\sqrt{d}}$, at least one of them will escape!

Main Lemma: measure the width



Escape Case

Lemma (Escape case)

Suppose we start with point $\tilde{x}$ satisfying following conditions:

$$\|\nabla f(\tilde{x})\| \leq g_{\text{thres}}, \qquad \lambda_{\min(\nabla^2 f(\tilde{x}))} \leq -\gamma$$

Let $x_0 = \tilde{x} + \xi$, where ξ come from the uniform distribution over ball with radius r, and let x_t be the iterates of GD from x_0. Then when $\eta < \frac{c_{\max}}{\ell}$, with at least probability $1 - \delta$, for any $T \geq t_{\text{thres}}$:

$$f(x_T) - f(\tilde{x}) \leq -f_{\text{thres}}$$

Proof of the escape lemma

Proof of the escape lemma

By smoothness, the perturbation step does not increase f much:

$$f(x_0) - f(\tilde{x}) \leq \nabla f(\tilde{x})^\top \xi + \frac{\ell}{2}\|\xi\|^2 \leq \cdots \leq 1.5 f_{\text{thres}}$$

Proof of the escape lemma

By smoothness, the perturbation step does not increase f much:

$$f(x_0) - f(\tilde{x}) \leq \nabla f(\tilde{x})^\top \xi + \frac{\ell}{2}\|\xi\|^2 \leq \cdots \leq 1.5 f_{\text{thres}}$$

By the main lemma, for any $x_0 \in \mathcal{X}_{\text{stuck}}$, we know $(x_0 \pm \mu r e_1) \notin \mathcal{X}_{\text{stuck}}$, where $\mu \in [\delta/(2\sqrt{d}), 1]$.

$$\text{Vol}(\mathcal{X}_{\text{stuck}}) = \text{Vol}(\mathbb{B}_{\tilde{x}}^{(d-1)}(r)) \times \frac{\delta r}{2\sqrt{d}} \times 2$$

Proof of the escape lemma

By smoothness, the perturbation step does not increase f much:

$$f(x_0) - f(\tilde{x}) \le \nabla f(\tilde{x})^\top \xi + \frac{\ell}{2}\|\xi\|^2 \le \cdots \le 1.5 f_{\text{thres}}$$

By the main lemma, for any $x_0 \in \mathcal{X}_{\text{stuck}}$, we know $(x_0 \pm \mu r e_1) \notin \mathcal{X}_{\text{stuck}}$, where $\mu \in [\delta/(2\sqrt{d}), 1]$.

$$\text{Vol}(\mathcal{X}_{\text{stuck}}) = \text{Vol}(\mathbb{B}_{\tilde{x}}^{(d-1)}(r)) \times \frac{\delta r}{2\sqrt{d}} \times 2$$

Therefore, the probability that we picked a point in $\mathcal{X}_{\text{stuck}}$ is bounded by

$$\frac{\text{Vol}(\mathcal{X}_{\text{stuck}})}{\text{Vol}(\mathbb{B}_{\tilde{x}}^{(d)}(r)))} \le \delta$$

Proof of the escape lemma

Thus, with probability at least $1 - \delta$, $x_0 \notin \mathcal{X}_{\text{stuck}}$, and in this case, by the main lemma.

$$f(x_T) - f(\tilde{x}) \leq -2.5 f_{\text{thres}} + 1.5 f_{\text{thres}} = -f_{\text{thres}}$$

How to prove the main Lemma?

How to prove the main Lemma?

- If u_T does not decrease function value, then $\{u_0, \cdots, u_T\}$ are close to $\tilde{x}$.

How to prove the main Lemma?

- If u_T does not decrease function value, then $\{u_0, \cdots, u_T\}$ are close to $\tilde{x}$.
- If $\{u_0, \cdots, u_T\}$ are close to $\tilde{x}$, GD on w_0 will decrease the function value.

How to prove the main Lemma?

- If u_T does not decrease function value, then $\{u_0, \cdots, u_T\}$ are close to $\tilde{x}$.
- If $\{u_0, \cdots, u_T\}$ are close to $\tilde{x}$, GD on w_0 will decrease the function value.

We will need the following approximation:

$$\tilde{f}_y(x) = f(y) + \nabla f(y)^\top (x - y) + \frac{1}{2}(x - y)^\top H(x - y)$$

where $H = \nabla^2 f(\tilde{x})$.

Two lemmas (simplified)

Lemma (u_T-stuck)

There exists absolute constant $c_{\max}$ *s.t., for any initial point* u_0 *with* $\|u_0 - \tilde{x}\| \le r$, *defined*

$$T = \min\left\{\inf_t \left\{t | \tilde{f}_{u_0}(u_t) - f(u_0) \le -3f_{\text{thres}}\right\}, t_{\text{thres}}\right\}$$

Then, for any $\eta \le \frac{c_{\max}}{\ell}$, *we have for all* $t < T$, $\|u_t - \tilde{x}\| \le \Phi$.

Two lemmas (simplified)

Lemma (u_T-stuck)

There exists absolute constant $c_{\max}$ s.t., for any initial point u_0 with $\|u_0 - \tilde{x}\| \le r$, defined

$$T = \min\left\{\inf_t \left\{t | \tilde{f}_{u_0}(u_t) - f(u_0) \le -3f_{\text{thres}}\right\}, t_{\text{thres}}\right\}$$

Then, for any $\eta \le \frac{c_{\max}}{\ell}$, we have for all $t < T$, $\|u_t - \tilde{x}\| \le \Phi$.

Lemma (w_T-escape)

There exists absolute constant $c_{\max}$ s.t., define

$$T = \min\left\{\inf_t \left\{t | \tilde{f}_{w_0}(w_t) - f(w_0) \le -3f_{\text{thres}}\right\}, t_{\text{thres}}\right\}$$

then, for any $\eta \le \frac{c_{\max}}{\ell}$, if $\|u_t - \tilde{x}\| \le \Phi$ for $t < T$, we have $T < t_{\text{thres}}$.

Prove the main lemma

Prove the main lemma

Assume $\tilde{x}$ is the origin. Define

$$T' = \inf_t \left\{ t | \tilde{f}_{u_0}(u_t) - f(u_0) \leq -3f_{\text{thres}} \right\}$$

Prove the main lemma

Assume $\tilde{x}$ is the origin. Define

$$T' = \inf_t \left\{ t | \tilde{f}_{u_0}(u_t) - f(u_0) \leq -3f_{\text{thres}} \right\}$$

Case $T' \leq t_{\text{thres}}$:
We know $\|u_{T'-1}\| \leq \Phi$ by u_T-stuck-lemma. By simple calculation, we can show that $\|u_{T'}\| = O(\Phi)$ as well.

Prove the main lemma

Assume $\tilde{x}$ is the origin. Define

$$T' = \inf_t \left\{ t | \tilde{f}_{u_0}(u_t) - f(u_0) \leq -3f_{\text{thres}} \right\}$$

Case $T' \leq t_{\text{thres}}$:
We know $\|u_{T'-1}\| \leq \Phi$ by u_T-stuck-lemma. By simple calculation, we can show that $\|u_{T'}\| = O(\Phi)$ as well.

$$\begin{aligned}
& f(u_{T'}) - f(u_0) \\
\leq & \nabla f(u_0)^\top (u_{T'} - u_0) + \frac{1}{2}(u_{T'} - u_0)^\top \nabla^2 f(u_0)(u_{T'} - u_0) + \frac{\rho}{6}\|u_{T'} - u_0\|^3 \\
\leq & \tilde{f}_{u_0}(u_t) - f(u_0) + \frac{\rho}{2}\|u_0 - \tilde{x}\|\|u_{T'} - u_0\|^2 + \frac{\rho}{6}\|u_{T'} - u_0\|^3 \\
\leq & -2.5f_{\text{thres}}
\end{aligned}$$

Prove the main lemma

Case $T' > t_{\text{thres}}$: By u_T-stuck-lemma, we know for all $t \leq t_{\text{thres}}$ $\|u_t\| \leq \Phi$.

Prove the main lemma

Case $T' > t_{\text{thres}}$: By u_T-stuck-lemma, we know for all $t \leq t_{\text{thres}}$ $\|u_t\| \leq \Phi$. Using the w_T-escape-lemma, we know

$$T'' = \inf_t \left\{ t | \tilde{f}_{w_0}(w_t) - f(w_0) \leq -3f_{\text{thres}} \right\} \leq t_{\text{thres}}$$

Prove the main lemma

Case $T' > t_{\text{thres}}$: By u_T-stuck-lemma, we know for all $t \leq t_{\text{thres}}$ $\|u_t\| \leq \Phi$. Using the w_T-escape-lemma, we know

$$T'' = \inf_t \left\{ t | \tilde{f}_{w_0}(w_t) - f(w_0) \leq -3f_{\text{thres}} \right\} \leq t_{\text{thres}}$$

Then we may reduce this to the case that $T' \leq t_{\text{thres}}$ because w, u are interchangeable.

Prove the u_T-stuck-lemma

Prove the u_T-stuck-lemma

Lemma (u_T-stuck)

There exists absolute constant $c_{\max}$ s.t., for any initial point u_0 with $\|u_0 - \tilde{x}\| \leq r$, defined

$$T = \min\left\{\inf_t \left\{t | \tilde{f}_{u_0}(u_t) - f(u_0) \leq -3f_{\text{thres}}\right\}, t_{\text{thres}}\right\}$$

Then, for any $\eta \leq \frac{c_{\max}}{\ell}$, we have for all $t < T$, $\|u_t - \tilde{x}\| \leq \Phi$.

Prove the u_T-stuck-lemma

Lemma (u_T-stuck)

There exists absolute constant $c_{\max}$ s.t., for any initial point u_0 with $\|u_0 - \tilde{x}\| \le r$, defined

$$T = \min\left\{\inf_t \left\{t | \tilde{f}_{u_0}(u_t) - f(u_0) \le -3f_{\text{thres}}\right\}, t_{\text{thres}}\right\}$$

Then, for any $\eta \le \frac{c_{\max}}{\ell}$, we have for all $t < T$, $\|u_t - \tilde{x}\| \le \Phi$.

- We won't move much in large negative eigenvector directions, otherwise it's a lot of progress!

Prove the u_T-stuck-lemma

Lemma (u_T-stuck)

There exists absolute constant $c_{\max}$ s.t., for any initial point u_0 with $\|u_0 - \tilde{x}\| \le r$, defined

$$T = \min\left\{\inf_t \left\{t | \tilde{f}_{u_0}(u_t) - f(u_0) \le -3f_{\text{thres}}\right\}, t_{\text{thres}}\right\}$$

Then, for any $\eta \le \frac{c_{\max}}{\ell}$, we have for all $t < T$, $\|u_t - \tilde{x}\| \le \Phi$.

- We won't move much in large negative eigenvector directions, otherwise it's a lot of progress!

Consider B_t as u_t in the remaining space where eigenvalue $\ge -\frac{\gamma}{100}$,

$$\|B_{t+1}\| \le (1 + \frac{\eta\gamma}{100})\|B_t\| + 2\eta g_{\text{thres}}$$

Prove the u_T-stuck-lemma

Lemma (u_T-stuck)

There exists absolute constant $c_{\max}$ s.t., for any initial point u_0 with $\|u_0 - \tilde{x}\| \le r$, defined

$$T = \min\left\{\inf_t\left\{t|\tilde{f}_{u_0}(u_t) - f(u_0) \le -3f_{\text{thres}}\right\}, t_{\text{thres}}\right\}$$

Then, for any $\eta \le \frac{c_{\max}}{\ell}$, we have for all $t < T$, $\|u_t - \tilde{x}\| \le \Phi$.

- We won't move much in large negative eigenvector directions, otherwise it's a lot of progress!

Consider B_t as u_t in the remaining space where eigenvalue $\ge -\frac{\gamma}{100}$,

$$\|B_{t+1}\| \le (1 + \frac{\eta\gamma}{100})\|B_t\| + 2\eta g_{\text{thres}}$$

If $T \le t_{\text{thres}}$, we will have $(1 + \frac{\eta\gamma}{100})^T \le 3$, so $\|B_T\|$ is bounded.

Prove the w_T-escape-lemma

Lemma (w_T-escape)

There exists absolute constant $c_{\max}$ *s.t., define*

$$T = \min\left\{\inf_t\left\{t|\tilde{f}_{w_0}(w_t) - f(w_0) \leq -3f_{\text{thres}}\right\}, t_{\text{thres}}\right\}$$

then, for any $\eta \leq \frac{c_{\max}}{\ell}$, *if* $\|u_t - \tilde{x}\| \leq \Phi$ *for* $t < T$, *we have* $T < t_{\text{thres}}$.

Prove the w_T-escape-lemma

- let $v_t = w_t - u_t$

Prove the w_T-escape-lemma

- let $v_t = w_t - u_t$
- We want to say for $T < t_{\text{thres}}$, w_T made progress.

Prove the w_T-escape-lemma

- let $v_t = w_t - u_t$
- We want to say for $T < t_{\text{thres}}$, w_T made progress.
- If w_t makes no progress, by u_T-stuck-lemma, it's still near $\tilde{x}$.

Prove the w_T-escape-lemma

- let $v_t = w_t - u_t$
- We want to say for $T < t_{\text{thres}}$, w_T made progress.
- If w_t makes no progress, by u_T-stuck-lemma, it's still near $\tilde{x}$.
- Therefore, we always have $\|v_t\| \leq \|u_t\| + \|w_t\| \leq 2\Phi$.

Prove the w_T-escape-lemma

- let $v_t = w_t - u_t$
- We want to say for $T < t_{\text{thres}}$, w_T made progress.
- If w_t makes no progress, by u_T-stuck-lemma, it's still near $\tilde{x}$.
- Therefore, we always have $\|v_t\| \leq \|u_t\| + \|w_t\| \leq 2\Phi$.

However, v_t is increasing very rapidly. It can't be always small!

Prove the w_T-escape-lemma

- let $v_t = w_t - u_t$
- We want to say for $T < t_{\text{thres}}$, w_T made progress.
- If w_t makes no progress, by u_T-stuck-lemma, it's still near $\tilde{x}$.
- Therefore, we always have $\|v_t\| \leq \|u_t\| + \|w_t\| \leq 2\Phi$.

However, v_t is increasing very rapidly. It can't be always small!

- At e_1 direction v_0 has at least $\frac{\delta r}{2\sqrt{d}}$

Prove the w_T-escape-lemma

- let $v_t = w_t - u_t$
- We want to say for $T < t_{\text{thres}}$, w_T made progress.
- If w_t makes no progress, by u_T-stuck-lemma, it's still near $\tilde{x}$.
- Therefore, we always have $\|v_t\| \leq \|u_t\| + \|w_t\| \leq 2\Phi$.

However, v_t is increasing very rapidly. It can't be always small!

- At e_1 direction v_0 has at least $\frac{\delta r}{2\sqrt{d}}$
- Every time it multiplies by at least $1 + \eta\gamma$.

Prove the w_T-escape-lemma

- let $v_t = w_t - u_t$
- We want to say for $T < t_{\text{thres}}$, w_T made progress.
- If w_t makes no progress, by u_T-stuck-lemma, it's still near $\tilde{x}$.
- Therefore, we always have $\|v_t\| \le \|u_t\| + \|w_t\| \le 2\Phi$.

However, v_t is increasing very rapidly. It can't be always small!

- At e_1 direction v_0 has at least $\frac{\delta r}{2\sqrt{d}}$
- Every time it multiplies by at least $1 + \eta\gamma$.
- In $T < t_{\text{thres}}$, we get $v_T > 2\Phi$, so w_T made progress!