Optimization for Machine Learning

Lecture 6: Tractable nonconvex problems

6.881: MIT

Suvrit Sra
Massachusetts Institute of Technology

04 Mar, 2021

Tractable nonconvex problems

Not all non-convex problems are bad

Tractable nonconvex problems

Not all non-convex problems are bad

- Generalizing the notion of convexity
- Problems with hidden convexity
- Miscellaneous examples from applications
- The list is much longer and growing!

Spectral problems

Simplest example: eigenvalues

Largest eigenvalue of a symmetric matrix

$$Ax = \lambda_{\max} x \quad \Leftrightarrow \quad \max_{x^T x = 1} \; x^T A x.$$

Nonconvex problem, but we know how to solve it!

Simplest example: eigenvalues

Largest eigenvalue of a symmetric matrix

$$Ax = \lambda_{\max} x \quad \Leftrightarrow \quad \max_{x^T x = 1} \; x^T A x.$$

Nonconvex problem, but we know how to solve it!

$$\begin{aligned} \mathcal{L}(x, \theta) := -x^T A x + \theta(x^T x - 1) \\ -2Ax + 2\theta x = 0 \\ Ax = \theta x \end{aligned}$$

Neccessary condition asks for (θ, x) to be eigenpair. Thus, $x^T A x$ is maximized by largest such pair.

Simplest example: eigenvalues

Largest eigenvalue of a symmetric matrix

$$Ax = \lambda_{\max} x \quad \Leftrightarrow \quad \max_{x^T x = 1} \; x^T A x.$$

Nonconvex problem, but we know how to solve it!

$$\begin{aligned} \mathcal{L}(x, \theta) := -x^T A x + \theta(x^T x - 1) \\ -2Ax + 2\theta x = 0 \\ Ax = \theta x \end{aligned}$$

Neccessary condition asks for (θ, x) to be eigenpair. Thus, $x^T A x$ is maximized by largest such pair. **Alternative:** Let $A = UDU^*$; then $\max_{x^T x = 1} x^T A x = \max_{y^T y = 1} \sum_i \lambda_i y_i^2$, where $y = U^* x$.

Simplest example: eigenvalues

Largest eigenvalue of a symmetric matrix

$$Ax = \lambda_{\max} x \quad \Leftrightarrow \quad \max_{x^T x = 1} \; x^T A x.$$

Nonconvex problem, but we know how to solve it!

$$\begin{aligned} \mathcal{L}(x, \theta) := -x^T A x + \theta(x^T x - 1) \\ -2Ax + 2\theta x = 0 \\ Ax = \theta x \end{aligned}$$

Neccessary condition asks for (θ, x) to be eigenpair. Thus, $x^T A x$ is maximized by largest such pair. **Alternative:** Let $A = UDU^*$; then $\max_{x^T x = 1} x^T A x = \max_{y^T y = 1} \sum_i \lambda_i y_i^2$, where $y = U^* x$.

$$\max_{y^T y = 1} \sum_i \lambda_i y_i^2 = \max_{z^T 1 = 1, z \geq 0} \sum\nolimits_i \lambda_i z_i,$$

which is a convex optimization problem.

Generalized eigenvalues

Let A, B be symmetric matrices; ***generalized eigenvalue*** is:

$$\max_{x \neq 0} \frac{x^T A x}{x^T B x}$$

(more generally: $Ax = \lambda Bx$, generalized eigenvectors)

Generalized eigenvalues

Let A, B be symmetric matrices; *generalized eigenvalue* is:

$$\max_{x \neq 0} \frac{x^T A x}{x^T B x}$$

(more generally: $Ax = \lambda Bx$, generalized eigenvectors)

Exercise: Study its Lagrangian formulation as well as a convex reformulation (similar to the "alternative" on slide 4)

Generalized eigenvalues

Let A, B be symmetric matrices; *generalized eigenvalue* is:

$$\max_{x \neq 0} \frac{x^T A x}{x^T B x}$$

(more generally: $Ax = \lambda Bx$, generalized eigenvectors)

Exercise: Study its Lagrangian formulation as well as a convex reformulation (similar to the "alternative" on slide 4)

Read the book: *https://web.stanford.edu/~boyd/lmibook/lmibook.pdf*

Trust region subproblem

$$\begin{aligned} \min_x \quad & x^T A x + 2b^T x + c \\ \text{s.t.} \quad & x^T B x + 2d^T x + e \leq 0. \end{aligned}$$

Here A and B are merely symmetric. Hence, nonconvex

Trust region subproblem

$$\begin{aligned} \min_{x} \quad & x^T A x + 2b^T x + c \\ \text{s.t.} \quad & x^T B x + 2d^T x + e \leq 0. \end{aligned}$$

Here A and B are merely symmetric. Hence, nonconvex

The dual problem can be formulated as (**Verify!**)

$$\begin{aligned} \max_{u,v\in\mathbb{R}} \quad & u \\ \text{s.t.} \quad & \begin{bmatrix} A + vB & b + vd \\ (b + vd)^T & c + ve - u \end{bmatrix} \succeq 0, \\ v \quad & \geq 0. \end{aligned}$$

Importantly, strong duality holds (see Appendix B of BV).
(alternatively: turns out SDP relaxation of the primal is exact)

Trust region subproblem

$$\begin{aligned} \min_{x} \quad & x^T A x + 2b^T x + c \\ \text{s.t.} \quad & x^T B x + 2d^T x + e \leq 0. \end{aligned}$$

Here A and B are merely symmetric. Hence, nonconvex

The dual problem can be formulated as (**Verify!**)

$$\begin{aligned} \max_{u,v \in \mathbb{R}} \quad & u \\ \text{s.t.} \quad & \begin{bmatrix} A + vB & b + vd \\ (b + vd)^T & c + ve - u \end{bmatrix} \succeq 0, \\ & v \geq 0. \end{aligned}$$

Importantly, strong duality holds (see Appendix B of BV).
(alternatively: turns out SDP relaxation of the primal is exact)

Ref: See Wang, Kılınç-Karzan, *The generalized trust-region subproblem: solution complexity and convex hull results*, 2019, for recent results.

Toeplitz-Hausdorff Theorem

Let A be a complex, square matrix. Its ***numerical range*** is

$$W(A) := \{x^*Ax \mid \|x\|_2 = 1, x \in \mathbb{C}^n\}.$$

Toeplitz-Hausdorff Theorem

Let A be a complex, square matrix. Its ***numerical range*** is

$$W(A) := \{x^* A x \mid \|x\|_2 = 1, x \in \mathbb{C}^n\}.$$

Theorem. The set $W(A)$ is convex (amazing!).

Toeplitz-Hausdorff Theorem

Let A be a complex, square matrix. Its ***numerical range*** is

$$W(A) := \{x^*Ax \mid \|x\|_2 = 1, x \in \mathbb{C}^n\}.$$

Theorem. The set $W(A)$ is convex (amazing!).

Exercise: If A is Hermitian show that $W(A) = [\lambda_{\min}, \lambda_{\max}]$.
Exercise: If $AA^* = A^*A$, then $W(A) = \operatorname{conv}(\lambda_i(A))$.

Explore: Let $A_1, \ldots, A_n$ be Hermitian. When is the set

$$\left\{(z^*A_1z, z^*A_2z, \ldots, z^*A_nz) \mid z \in \mathbb{C}^d, \|z\| = 1\right\}$$

convex (this is also called the "*joint numerical range*").

Principal Component Analysis (PCA)

Let $A \in \mathbb{R}^{n \times p}$. Consider the nonconvex problem

$$\min_{X} \quad \|A - X\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \operatorname{rank}(X) = k.$$

Principal Component Analysis (PCA)

Let $A \in \mathbb{R}^{n \times p}$. Consider the nonconvex problem

$$\min_X \quad \|A - X\|_F^2 \quad \text{s.t.} \quad \operatorname{rank}(X) = k.$$

Well-known Eckart-Young-Mirsky theorem shows that

$$X^* = U_k \Sigma_k V_k^T$$

where A has the SVD $A = U \Sigma V^T$.

Why is this true?

PCA via the Fantope

Another characterization of SVD (nonconvex prob)

$$\min_{Z=Z^T} \|A - AZ\|_F^2, \qquad \text{s.t.} \quad \operatorname{rank}(Z) = k, Z \text{ is a projection}$$

$$\Leftrightarrow \max_{Z=Z^T} \langle A^T A,\ Z\rangle, \qquad \text{s.t.} \quad \operatorname{rank}(Z) = k, Z \text{ is a projection.}$$

PCA via the Fantope

Another characterization of SVD (nonconvex prob)

$$\min_{Z=Z^T} \|A - AZ\|_F^2, \qquad \text{s.t.} \quad \text{rank}(Z) = k, Z \text{ is a projection}$$
$$\Leftrightarrow \max_{Z=Z^T} \langle A^T A,\ Z\rangle, \qquad \text{s.t.} \quad \text{rank}(Z) = k, Z \text{ is a projection.}$$

Optimal solution here is $Z = V_k V_k^T$, the top-k evecs of $A^T A$

PCA via the Fantope

Another characterization of SVD (nonconvex prob)

$$\min_{Z=Z^T} \|A - AZ\|_F^2, \qquad \text{s.t.} \quad \text{rank}(Z) = k, Z \text{ is a projection}$$

$$\Leftrightarrow \max_{Z=Z^T} \langle A^T A,\ Z \rangle, \qquad \text{s.t.} \quad \text{rank}(Z) = k, Z \text{ is a projection.}$$

Optimal solution here is $Z = V_k V_k^T$, the top-k evecs of $A^T A$

Equivalent convex problem!

PCA via the Fantope

Another characterization of SVD (nonconvex prob)

$$\min_{Z=Z^T} \|A - AZ\|_F^2, \quad \text{s.t.} \quad \mathrm{rank}(Z) = k, Z \text{ is a projection}$$
$$\Leftrightarrow \max_{Z=Z^T} \langle A^T A,\ Z\rangle, \quad \text{s.t.} \quad \mathrm{rank}(Z) = k, Z \text{ is a projection}.$$

Optimal solution here is $Z = V_k V_k^T$, the top-k evecs of $A^T A$

Equivalent convex problem!

First, write constraint set C as

$$C = \{Z = Z^T \mid \mathrm{rank}(Z) = k, Z \text{ is a projection}\}$$

PCA via the Fantope

Another characterization of SVD (nonconvex prob)

$$\min_{Z=Z^T} \|A - AZ\|_F^2, \quad \text{s.t.} \quad \mathrm{rank}(Z) = k, Z \text{ is a projection}$$
$$\Leftrightarrow \max_{Z=Z^T} \langle A^T A,\ Z\rangle, \quad \text{s.t.} \quad \mathrm{rank}(Z) = k, Z \text{ is a projection.}$$

Optimal solution here is $Z = V_k V_k^T$, the top-k evecs of $A^T A$

Equivalent convex problem!

First, write constraint set C as

$$\begin{aligned} C &= \{Z = Z^T \mid \mathrm{rank}(Z) = k, Z \text{ is a projection}\} \\ &= \{Z = Z^T \mid \lambda_i(Z) \in \{0,1\}, \mathrm{Tr}(Z) = k\}. \end{aligned}$$

Fantope

Now consider convex hull: $\mathcal{C} = \text{conv}\, C$

Fantope

Now consider convex hull: $\mathcal{C} = \text{conv}\ C$

$$\mathcal{C} \quad = \quad \left\{ Z = Z^T \mid \lambda_i(Z) \in [0,1], \operatorname{Tr}(Z) = k \right\}$$

Fantope

Now consider convex hull: $\mathcal{C} = \text{conv}\, C$

$$
\begin{aligned}
\mathcal{C} &= \left\{ Z = Z^T \mid \lambda_i(Z) \in [0,1], \operatorname{Tr}(Z) = k \right\} \\
&= \left\{ Z = Z^T \mid 0 \preceq Z \preceq I, \operatorname{Tr}(Z) = k \right\}.
\end{aligned}
$$

The set $\mathcal{C}$ is called the ***Fantope*** (named after Ky Fan).

Fantope

Now consider convex hull: $\mathcal{C} = \text{conv } C$

$$\begin{aligned} \mathcal{C} &= \left\{ Z = Z^T \mid \lambda_i(Z) \in [0,1], \operatorname{Tr}(Z) = k \right\} \\ &= \left\{ Z = Z^T \mid 0 \preceq Z \preceq I, \operatorname{Tr}(Z) = k \right\}. \end{aligned}$$

The set $\mathcal{C}$ is called the ***Fantope*** (named after Ky Fan).

Exercise: Now invoke the "maximize a convex function" idea from Lecture 5 to claim that the convex problem $\max_{Z=Z^T} \langle A^T A, Z \rangle$ s.t. $Z \in \mathcal{C}$ solves the original problem.

Sparsity

Nonconvex Sparse optimization

The ℓ_0-quasi-norm is defined as

$$\|x\|_0 := \text{card}\{x_i \mid x_i \neq 0\}.$$

Nonconvex Sparse optimization

The ℓ_0-quasi-norm is defined as

$$\|x\|_0 := \operatorname{card}\{x_i \mid x_i \neq 0\}.$$

Projection onto ℓ_0-ball

$$\min \quad \tfrac{1}{2}\|x - y\|_2^2, \quad \text{s.t.} \quad \|x\|_0 \leq k.$$

Nonconvex Sparse optimization

The ℓ_0-quasi-norm is defined as

$$\|x\|_0 := \operatorname{card}\{x_i \mid x_i \neq 0\}.$$

Projection onto ℓ_0-ball

$$\min \quad \tfrac{1}{2}\|x-y\|_2^2, \quad \text{s.t.} \quad \|x\|_0 \leq k.$$

Nonconvex but tractable: If $\|y\|_0 \leq k$, then clearly $x = y$.
Otherwise, pick the k largest entries of $|y|$, and set the rest to 0.

Nonconvex Sparse optimization

The ℓ_0-quasi-norm is defined as

$$\|x\|_0 := \operatorname{card}\{x_i \mid x_i \neq 0\}.$$

Projection onto ℓ_0-ball

$$\min \quad \tfrac{1}{2}\|x-y\|_2^2, \quad \text{s.t.} \quad \|x\|_0 \leq k.$$

Nonconvex but tractable: If $\|y\|_0 \leq k$, then clearly $x = y$.
Otherwise, pick the k largest entries of $|y|$, and set the rest to 0.

Exercise: Prove the above claim.

Exercise: Similarly solve $\frac{1}{2}\|x-y\|_2^2 + \lambda\|x\|_0$

Used in so-called "Iterative Hard Thresholding" algorithms

Compressed Sensing

$$\min \quad \|x\|_0 \quad \text{s.t.} \quad Ax = b$$

Compressed Sensing

$$\min \quad \|x\|_0 \quad \text{s.t.} \quad Ax = b$$

If the "measurement matrix" A satisfies so-called *restricted isometry condition* with the constant $\delta_s \in (0, 1)$

$$(1 - \delta_s)\|x\|^2 \leq \|Ax\|^2 \leq (1 + \delta_s)\|x\|^2, \qquad x \text{ is } s\text{-sparse},$$

then the ℓ_1-convex relaxation is exact.

Explore: (search keywords): compressed sensing, sparse recovery, restricted isometry

Generalized convexity

Geometric programming

Monomial: $g : \mathbb{R}^n_{++} \to \mathbb{R}$ of the form

$$g(x) = \gamma x_1^{a_1} \cdots x_n^{a_n}, \quad \gamma > 0, a_i \in \mathbb{R}.$$

Posynomial: Sum of monomials, e.g, $f(x) = \sum_j g_j(x)$

Geometric programming

Monomial: $g : \mathbb{R}^n_{++} \to \mathbb{R}$ of the form

$$g(x) = \gamma x_1^{a_1} \cdots x_n^{a_n}, \quad \gamma > 0, a_i \in \mathbb{R}.$$

Posynomial: Sum of monomials, e.g, $f(x) = \sum_j g_j(x)$

Geometric Program

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & f_i(x) \leq 1, \quad i \in [m] \\ & g_j(x) = 1, \quad j \in [r], \end{aligned}$$

where f_i are posynomials and g_j are monomials.

Clearly, **nonconvex**.

Geometric programming

Make change of variables: $y_i = \log x_i$ (recall $x_i > 0$). Then,

$$f(x) = f(e^y) = \gamma(e^{y_1})^{a_1} \cdots (e^{y_n})^{a_n} = e^{a^T y + b},$$

for $b = \log y$.

Geometric programming

Make change of variables: $y_i = \log x_i$ (recall $x_i > 0$). Then,

$$f(x) = f(e^y) = \gamma(e^{y_1})^{a_1} \cdots (e^{y_n})^{a_n} = e^{a^T y + b},$$

for $b = \log \gamma$. Thus, after taking logs, *geometric program* is

$$\begin{aligned} \min_y \quad & \log\left(\sum\nolimits_k e^{a_{0k}^T y + b_{0k}}\right) \\ \text{s.t.} \quad & \log\left(\sum\nolimits_k e^{a_{0k}^T y + b_{0k}}\right) \leq 0, i \in [m] \\ & c_j^T y + d_j = 0, j \in [r], \end{aligned}$$

for suitable sets of vectors $\{a_{ik}\}$, and $\{c_j\}$.

Geometric programming

Make change of variables: $y_i = \log x_i$ (recall $x_i > 0$). Then,

$$f(x) = f(e^y) = \gamma(e^{y_1})^{a_1} \cdots (e^{y_n})^{a_n} = e^{a^T y + b},$$

for $b = \log y$. Thus, after taking logs, *geometric program* is

$$\begin{aligned} \min_y \quad & \log\left(\sum\nolimits_k e^{a_{0k}^T y + b_{0k}}\right) \\ \text{s.t.} \quad & \log\left(\sum\nolimits_k e^{a_{0k}^T y + b_{0k}}\right) \leq 0, i \in [m] \\ & c_j^T y + d_j = 0, j \in [r], \end{aligned}$$

for suitable sets of vectors $\{a_{ik}\}$, and $\{c_j\}$.
Recall, log-sum-exp is convex, so above is a convex opt.

Ref: See Chapter 8.8 of BV; search online for "geometric programming"

Generalized convexity

- Quasiconvexity: If level sets $L_t(f) = \{x \mid f(x) \leq t\}$ are convex, we say f is ***quasiconvex***

Generalized convexity

- Quasiconvexity: If level sets $L_t(f) = \{x \mid f(x) \leq t\}$ are convex, we say f is ***quasiconvex***
- Arcwise Convexity: $f(\gamma_{xy}(t)) \leq (1-t)f(x) + tf(y)$, where ***arc*** $\gamma : [0,1] \to X$ joins point x to point y.

Generalized convexity

- Quasiconvexity: If level sets $L_t(f) = \{x \mid f(x) \le t\}$ are convex, we say f is ***quasiconvex***
- Arcwise Convexity: $f(\gamma_{xy}(t)) \le (1-t)f(x) + tf(y)$, where ***arc*** $\gamma : [0,1] \to X$ joins point x to point y.
- Several other notions of generalized convexity exist (see also: genconv.org!)

Generalized convexity

- Quasiconvexity: If level sets $L_t(f) = \{x \mid f(x) \leq t\}$ are convex, we say f is ***quasiconvex***
- Arcwise Convexity: $f(\gamma_{xy}(t)) \leq (1-t)f(x) + tf(y)$, where ***arc*** $\gamma : [0,1] \to X$ joins point x to point y.
- Several other notions of generalized convexity exist (see also: genconv.org!)

Exercise: Suppose a set X is arcwise convex, and $f : X \to \mathbb{R}$ is an arcwise convex function. Prove that a local optimum of f is also global (assume regularity as needed).

Exercise: View GP as arcwise convexity using: $\gamma(t) = x^{1-t}y^t$

Linear fractional programming

$$
\begin{aligned}
\min \quad & \frac{a^T x + b}{c^T x + d} \\
\text{s.t.} \quad & Gx \leq h, c^T x + d > 0, Ex = f.
\end{aligned}
$$

This problem is nonconvex, but it is quasiconvex.

Linear fractional programming

$$\begin{array}{ll} \min & \dfrac{a^T x + b}{c^T x + d} \\ \text{s.t.} & Gx \leq h, c^T x + d > 0, Ex = f. \end{array}$$

This problem is nonconvex, but it is quasiconvex. Provided it is feasible, it is equivalent to the LP

$$\begin{array}{ll} \min_{y,z} & a^T y + bz \\ \text{s.t.} & Gy - hz \leq 0, z \geq 0 \\ & Ey = fz, c^T y + dz = 1. \end{array}$$

Linear fractional programming

$$\begin{array}{ll} \min & \dfrac{a^T x + b}{c^T x + d} \\ \text{s.t.} & Gx \leq h, c^T x + d > 0, Ex = f. \end{array}$$

This problem is nonconvex, but it is quasiconvex. Provided it is feasible, it is equivalent to the LP

$$\begin{array}{ll} \min_{y,z} & a^T y + bz \\ \text{s.t.} & Gy - hz \leq 0, z \geq 0 \\ & Ey = fz, c^T y + dz = 1. \end{array}$$

These two problems connected via the transformation

$$y = \frac{x}{c^T x + d}, \quad z = \frac{1}{c^T x + d}.$$

See BV Chapter 4 for details.

Generalized Perron-Frobenius

Let $A, B \in \mathbb{R}^{m\times n}$.

$$\begin{aligned} \max_{x,\lambda} \quad & \lambda \\ \text{s.t.} \quad & \lambda A x \le Bx, x^T 1 = 1, x \ge 0. \end{aligned}$$

Exercise: Try solving it directly somehow.

Exercise: Cast this as an (extended) linear-fractional program.

Challenge: Simplex convexity

Let Δ_n be the probability simplex, i.e., set of vectors $x = (x_1, \ldots, x_n)$ such that $x_i \geq 0$ and $x^T 1 = 1$. Assume that $n \geq 2$. Prove that the following "Bethe entropy"

$$g(x) = \sum\nolimits_i x_i \log \frac{1}{x_i} + (1 - x_i) \log(1 - x_i),$$

is concave on Δ_n.

The Polyak-Łojasiewicz class

PL class aka gradient-dominated

$$f(x) - f(x^*) \le \tau \|\nabla f(x)\|^{\alpha}, \quad \alpha \ge 1.$$

Observe that if $\nabla f(x) = 0$, then x must be global opt.

The Polyak-Łojasiewicz class

PL class aka gradient-dominated

$$f(x) - f(x^*) \leq \tau \|\nabla f(x)\|^\alpha, \quad \alpha \geq 1.$$

Observe that if $\nabla f(x) = 0$, then x must be global opt.

Exercise: Let f be convex on $\mathbb{R}^n$. Prove that on the set $\{x \mid \|x - x^*\| \leq R\}$, f is PL with $\tau = R$ and $\alpha = 1$.

The Polyak-Łojasiewicz class

PL class aka gradient-dominated

$$f(x) - f(x^*) \leq \tau \|\nabla f(x)\|^\alpha, \quad \alpha \geq 1.$$

Observe that if $\nabla f(x) = 0$, then x must be global opt.

Exercise: Let f be convex on $\mathbb{R}^n$. Prove that on the set $\{x \mid \|x - x^*\| \leq R\}$, f is PL with $\tau = R$ and $\alpha = 1$.

Exercise: Let f be strongly-convex with parameter μ. Prove that f is a PL function with $\tau = 1/2\mu$ and $\alpha = 2$.

Important non-convex PL example

- Let $g(x) = (g_1(x), \ldots, g_m(x))$ be a differentiable func.

Important non-convex PL example

- Let $g(x) = (g_1(x), \ldots, g_m(x))$ be a differentiable func.
- Consider the system of nonlinear equations $g(x) = 0$

Important non-convex PL example

- Let $g(x) = (g_1(x), \ldots, g_m(x))$ be a differentiable func.
- Consider the system of nonlinear equations $g(x) = 0$
- Assume that $m \leq n$ and that $\exists x^*$ s.t. $g(x^*) = 0$.

Important non-convex PL example

- Let $g(x) = (g_1(x), \dots, g_m(x))$ be a differentiable func.
- Consider the system of nonlinear equations $g(x) = 0$
- Assume that $m \leq n$ and that $\exists x^*$ s.t. $g(x^*) = 0$.
- Assume Jacobian $J(x) = (\nabla g_1(x), \dots, \nabla g_m(x))$ non-degenerate on a convex set $\mathcal{X}$ containing x^*. Then, $\sigma = \inf_{x \in \mathcal{X}} \lambda_{min}(J(x)^T J(x)) > 0$.

Important non-convex PL example

- Let $g(x) = (g_1(x), \ldots, g_m(x))$ be a differentiable func.
- Consider the system of nonlinear equations $g(x) = 0$
- Assume that $m \leq n$ and that $\exists x^*$ s.t. $g(x^*) = 0$.
- Assume Jacobian $J(x) = (\nabla g_1(x), \ldots, \nabla g_m(x))$ non-degenerate on a convex set $\mathcal{X}$ containing x^*. Then, $\sigma = \inf_{x \in \mathcal{X}} \lambda_{min}(J(x)^T J(x)) > 0$.
- Let $f(x) = \frac{1}{2}\sum_i g_i^2(x)$; note that $\nabla f(x) = J(x)g(x)$

Important non-convex PL example

- Let $g(x) = (g_1(x), \ldots, g_m(x))$ be a differentiable func.
- Consider the system of nonlinear equations $g(x) = 0$
- Assume that $m \leq n$ and that $\exists x^*$ s.t. $g(x^*) = 0$.
- Assume Jacobian $J(x) = (\nabla g_1(x), \ldots, \nabla g_m(x))$ non-degenerate on a convex set $\mathcal{X}$ containing x^*. Then, $\sigma = \inf_{x \in \mathcal{X}} \lambda_{min}(J(x)^T J(x)) > 0$.
- Let $f(x) = \frac{1}{2}\sum_i g_i^2(x)$; note that $\nabla f(x) = J(x)g(x)$

$$\|\nabla f(x)\|^2 = g(x)^T J(x)^T J(x) g(x) \geq \sigma \|g(x)\|^2 = 2\sigma(f(x) - f(x^*))$$

Thus, f is PL with $\tau = 1/2\sigma$, $\alpha = 2$.

Important non-convex PL example

- Let $g(x) = (g_1(x), \ldots, g_m(x))$ be a differentiable func.
- Consider the system of nonlinear equations $g(x) = 0$
- Assume that $m \le n$ and that $\exists x^*$ s.t. $g(x^*) = 0$.
- Assume Jacobian $J(x) = (\nabla g_1(x), \ldots, \nabla g_m(x))$ non-degenerate on a convex set $\mathcal{X}$ containing x^*. Then, $\sigma = \inf_{x\in\mathcal{X}} \lambda_{min}(J(x)^T J(x)) > 0$.
- Let $f(x) = \frac{1}{2}\sum_i g_i^2(x)$; note that $\nabla f(x) = J(x)g(x)$

$$\|\nabla f(x)\|^2 = g(x)^T J(x)^T J(x) g(x) \ge \sigma \|g(x)\|^2 = 2\sigma(f(x) - f(x^*))$$

Thus, f is PL with $\tau = 1/2\sigma$, $\alpha = 2$.

Exercise: When $m < n$, are the Hessians of f degenerate at solutions?

Important non-convex PL example

- Let $g(x) = (g_1(x), \ldots, g_m(x))$ be a differentiable func.
- Consider the system of nonlinear equations $g(x) = 0$
- Assume that $m \leq n$ and that $\exists x^*$ s.t. $g(x^*) = 0$.
- Assume Jacobian $J(x) = (\nabla g_1(x), \ldots, \nabla g_m(x))$ non-degenerate on a convex set $\mathcal{X}$ containing x^*. Then, $\sigma = \inf_{x \in \mathcal{X}} \lambda_{min}(J(x)^T J(x)) > 0$.
- Let $f(x) = \frac{1}{2}\sum_i g_i^2(x)$; note that $\nabla f(x) = J(x)g(x)$

$$\|\nabla f(x)\|^2 = g(x)^T J(x)^T J(x) g(x) \geq \sigma \|g(x)\|^2 = 2\sigma(f(x) - f(x^*))$$

Thus, f is PL with $\tau = 1/2\sigma$, $\alpha = 2$.

Exercise: When $m < n$, are the Hessians of f degenerate at solutions?
Explore: Hamed Karimi, Julie Nutini, and Mark Schmidt. *Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition.* *https://arxiv.org/abs/1608.04636*

Others tractable nonconvex problems

- Instances of matrix completion, deep linear neural networks, tensor factorization, many others. Check out the great collection by Ju Sun: *https://sunju.org/research/nonconvex/*

Others tractable nonconvex problems

- Instances of matrix completion, deep linear neural networks, tensor factorization, many others. Check out the great collection by Ju Sun: *https://sunju.org/research/nonconvex/*
- Submodular optimization (later in course)
- Any combinatorial problem whose convex relaxation is tight

Others tractable nonconvex problems

- Instances of matrix completion, deep linear neural networks, tensor factorization, many others. Check out the great collection by Ju Sun: *https://sunju.org/research/nonconvex/*
- Submodular optimization (later in course)
- Any combinatorial problem whose convex relaxation is tight
- Non-Eucidean convexity (hinted at today, later in course)

Others tractable nonconvex problems

- Instances of matrix completion, deep linear neural networks, tensor factorization, many others. Check out the great collection by Ju Sun: *https://sunju.org/research/nonconvex/*
- Submodular optimization (later in course)
- Any combinatorial problem whose convex relaxation is tight
- Non-Eucidean convexity (hinted at today, later in course)

Example without "spurious" local minima: Deep Linear Network

Others tractable nonconvex problems

- Instances of matrix completion, deep linear neural networks, tensor factorization, many others. Check out the great collection by Ju Sun: *https://sunju.org/research/nonconvex/*
- Submodular optimization (later in course)
- Any combinatorial problem whose convex relaxation is tight
- Non-Eucidean convexity (hinted at today, later in course)

Example without "spurious" local minima: Deep Linear Network

$$\min \; L(W_1, \ldots, W_L) = \tfrac{1}{2}\|W_L W_{L-1} \cdots W_1 X - Y\|_F^2,$$

here $X \in \mathbb{R}^{d_x \times n}$: data/input matrix; and $Y \in \mathbb{R}^{d_y \times n}$ "label"/output matrix.

Others tractable nonconvex problems

- Instances of matrix completion, deep linear neural networks, tensor factorization, many others. Check out the great collection by Ju Sun: *https://sunju.org/research/nonconvex/*
- Submodular optimization (later in course)
- Any combinatorial problem whose convex relaxation is tight
- Non-Eucidean convexity (hinted at today, later in course)

Example without "spurious" local minima: Deep Linear Network

$$\min \; L(W_1, \ldots, W_L) = \tfrac{1}{2}\|W_L W_{L-1} \cdots W_1 X - Y\|_F^2,$$

here $X \in \mathbb{R}^{d_x \times n}$: data/input matrix; and $Y \in \mathbb{R}^{d_y \times n}$ "label"/output matrix.

Theorem. Let $k = \min(d_x, d_y)$ be the "width" of the network. Let $V = \{(W_1, \ldots, W_L) \mid \operatorname{rank}(\prod_l W_l) = k\}$. Then, every critical point of $L(W)$ in V is a global minimum, while every critical point in V^c is a saddle point.

Ref. Chulhee Yun, Suvrit Sra, Ali Jadbabaie. *Global optimality conditions for deep neural networks*. ICLR 2018.