Nonconvex Optimization at Large

min \( f(x) \)
subject to \( x \in D \).

“easy”

“hard”

Nonconvex Optimization Problems

In theory: NP-hard even to compute a local minimizer
In practice: Simple methods such as gradient descent
and alternating minimization often highly successful

Which Nonconvex Optimization Problems are “Easy”?

We will focus on the \( \chi \) family:
1. \( \chi \) (P-1) All local minimizers are also global
2. \( \chi \) (P-2) Around any saddle point there is a negative directional curvature (Ridable saddles)

Not all saddles are ridable (strict) – i.e., with indefinite Hessian!

One example from the family...

Quantitative Version of (P-2): Ridable (Strict)-Saddle Functions

A function \( f \) over manifold \( M \) is \( (\alpha, \beta, \gamma, \delta) \)-ridable \( (\alpha, \beta, \gamma, \delta \) strictly positive) if any \( x \in M \) obeys at least one of the following: \( f(T_x M) \) is the tangent space of \( M \) at point \( x \):
1. \( \text{[Strong gradient]} \ |\| \text{grad} f(x) \| \| \geq \beta \)
2. \( \text{[Negative curvature]} \) There exists \( v \in T_x M \) with \( \| v \| = 1 \) such that \( \langle \text{Hess}(f)(x), v \rangle \leq -\alpha \);
3. \( \text{[Strong convexity around minimizers]} \) There exists a local minimizer \( x \), such that \( \| x - x^\ast \| \leq \delta \), and for all \( y \in M \) that is in \( \delta \)-neighborhood of \( x \), \( \langle \text{Hess}(f)(y), v \rangle \geq \gamma \) for any \( v \in T_y M \) with \( \| v \| = 1 \), i.e., the function \( f \) is \( \gamma \)-strongly convex in \( \delta \)-neighborhood of \( x \).

Example I: Eigenvector Problem [Classic]

For a symmetric matrix \( A \in \mathbb{R}^{n \times n} \),

\[
\max_{x \in \mathbb{R}^n} x^T A x \text{ subject to } \| x \| = 1.
\]

- Critical points: \( \pm v \).
- Suppose \( \lambda_1 > \lambda_2 \geq \ldots \lambda_{n-1} > \lambda_n \). The only local/global minimizers are \( \pm v \), and all \( \pm v \) for \( 2 \leq i < n-1 \) are ridable saddles.
- Quantitatively, the function is \( (\epsilon(\lambda_{n-1} - \lambda_0), \epsilon(\lambda_{n-1} - \lambda_0)/\lambda_0, \epsilon(\lambda_{n-1} - \lambda_0), 2\epsilon(\lambda_{n-1} - \lambda_0)/\lambda_0) \)-ridable over \( \mathbb{S}^{n-1} \).

Example II: Sparse (Complete) Dictionary Learning [Sun et al’15]

Given \( Y \), find \( (A, X) \) such that \( Y \approx AX \), with \( X \) as sparse as possible.
When \( A \) square, invertible, \( \text{rank}(Y) = \text{rank}(X) \) --- Finding sparse vectors in \( \text{row}(Y) \)

\[
\min_{\| q \|_0} \sum_{i=1}^m h(q_t \beta_t) \text{ subject to } \| q \|_0 = 1. \quad (h \text{ promotes sparsity})
\]

Under Bernoulli-Gaussian model for \( X \) and large \( \rho \): * Any local minimizer recovers a row of \( X \) * \( f \) is \( (\epsilon, \epsilon, \epsilon, \mu, 2\sqrt{\epsilon}/\mu) \)-ridable over \( \mathbb{S}^{n-1} \).

Example III: Generalized Phase Retrieval [Sun et al’16]

Recover \( x \) from nonlinear measurements \( y_i = |a_i x|, k = 1, \ldots, m \), encountered in optic imaging and others

\[
\min_{\| q \|_0} \sum_{i=1}^m \beta(q_t \beta_t) \text{ subject to } \| q \|_0 = 1. \quad (h \text{ promotes sparsity})
\]

When \( a_i \)’s are Gaussian and \( m \geq O(n \log^2 n) \):
- All local minimizers are \( x \) or its equivalent copies
- The function \( f \) is \( (\epsilon, \epsilon, \epsilon, \epsilon, \mu) \)-ridable (modulo the “flat” direction)

Example IV: Tensor Decomposition and ICA [Ge et al’15]

Given orthogonally decomposable \( t \)-th order tensors \( T \), i.e., \( T = \sum_{i,j} a_{i,j} \), find \( a_i \)’s up to sign and permutation.

\[
\min_{a_i,j} \sum_{i,j} \| T_{i,j} - a_i a_j \|_2^2 = \sum_{i,j} \| T_{i,j} - a_i a_j \|_2^2, \quad \text{subject to } \| a_i \| = 1 \forall i \in [n].
\]

All local minimizers of \( g \) are equivalent (i.e., signed permuted) copies of \( \{a_1, \ldots, a_n\} \). Moreover, \( g \) is \( O(1/\text{poly}(n), 1/\text{poly}(n), 1/\text{poly}(n)) \)-ridable.

Algorithm: Second-order Trust-region Method (Several other possibilities)

Iteratively construct local quadratic approximations:

\[
\hat{f}(x^{(k)}) = f(x^{(k)}) + \frac{1}{2} (x - x^{(k)})^T \text{grad} f(x^{(k)}) + \frac{1}{2} \text{Hess}(f)(x^{(k)})(x - x^{(k)})
\]

The next movement determined by minimizing the quadratic approximation within a small radius \( \delta \):

\[
\delta^{(k+1)} = \arg \min \{ |T_{a_p a_q}(x^{(k)}) + \text{grad} f(x^{(k)T})| \} \|
\]

The next iterate is obtained by retracting the resulting vector back to the manifold:

\[
\hat{x}^{(k+1)} = R_{\text{grad} f(x^{(k)})} \delta^{(k+1)}.
\]

References

Sun et al. When Are Nonconvex Problems Not Scary?, arXiv:1510.08096
Boumal N. Nonconvex Phase Synchronization, arXiv:1601.06114

Email: cjs4038, qq2105, jm2966@columbia.edu