Escaping Saddle Points in Constrained Optimization

Massachusetts Institute of Technology

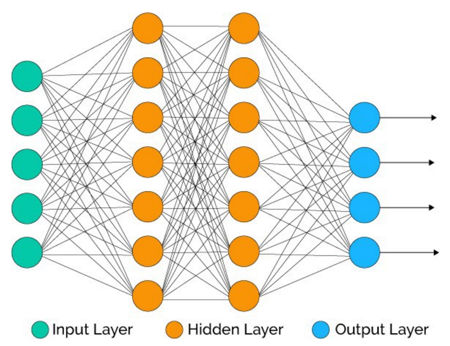
Aryan Mokhtari, Asuman Ozdaglar, and Ali Jadbabaie

Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology (MIT)

LIDS

IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY

Introduction



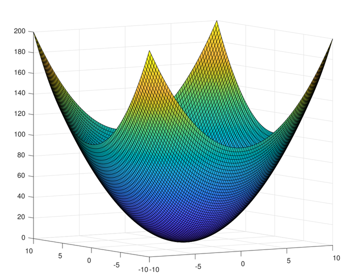
- Recent revival of interest in nonconvex optimization
  - $\Rightarrow$ Practical success and advances in computational tools
- Consider the following general optimization program

$$\min_{x \in \mathcal{C}} f(x)$$

- $\mathcal{C} \subseteq \mathbb{R}^d$ is a convex compact closed set $\Rightarrow$ This problem is hard

Convex Optimization: Optimality Condition

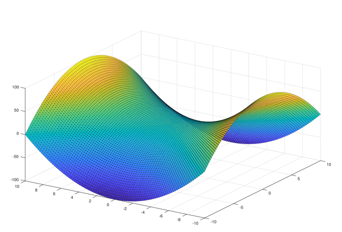
- Before jumping to nonconvex optimization
  - $\Rightarrow$ Let's recap the convex case!

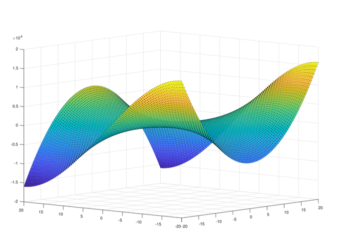


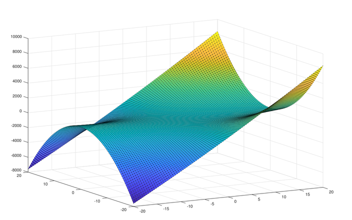
- In the convex setting (f is convex)
  - $\Rightarrow$ First-order optimality condition implies global optimality
  - $\Rightarrow$ Finding an approximate first-order stationary point is suff.

$$\begin{cases} \text{Unconstrained: Find } x^* \text{ s.t. } & \|\nabla f(x^*)\| \leq \varepsilon \\ \text{Constrained: Find } x^* \text{ s.t. } & \nabla f(x^*)^T(x - x^*) \geq -\varepsilon \quad \text{for all } x \in \mathcal{C} \end{cases}$$

Nonconvex Optimization







- 1st-order optimality is not enough $\Rightarrow$ Saddle points exist!
- Check higher order derivatives $\Rightarrow$ To escape from saddle points
  - $\Rightarrow$ Search for a second-order stationary point (SOSP)
- Does convergence to an SOSP lead to global optimality? No!
- But, if all saddles are escapable (strict saddles)
  - $\Rightarrow$ SOSP $\Rightarrow$ local minimum!
- In several cases, all saddle points are escapable and all local minima are global
  - $\Rightarrow$ Eigenvector problem [Absil et al., '10]
  - $\Rightarrow$ Phase retrieval [Sun et al., '16]
  - $\Rightarrow$ Dictionary learning [Sun et al., '17]

Unconstrained Optimization

- Consider the unconstrained nonconvex setting ($\mathcal{C} = \mathbb{R}^d$)
- x^* is an approximate (ε, γ)-second-order stationary point if

$$\underbrace{\|\nabla f(x^*)\| \leq \varepsilon}_{\text{first-order optimality condition}} \quad \text{and} \quad \underbrace{\nabla^2 f(x^*) \succeq -\gamma \mathbf{I}}_{\text{second-order optimality condition}}$$

- Various attempts to design algorithms converging to an SOSP
- Perturbing iterates by injecting noise
  - $\Rightarrow$ [Ge et al., '15], [Jin et al., '17a,b], [Daneshmand et al., '18]
- Using the eigenvector of the smallest eigenvalue of the Hessian
  - $\Rightarrow$ [Carmon et al., '16], [Allen-Zhu, '17], [Xu & Yang, '17], [Royer & Wright, '17], [Agarwal et al., '17], [Reddi et al., '18]
- Overall cost to find an (ε, γ)-SOSP $\Rightarrow$ Polynomial in ε^{-1} and γ^{-1}
- However, not applicable to the convex constrained setting!
- In the constrained case, can we find an SOSP in poly-time?

Constrained optimization: Second-order stationary point

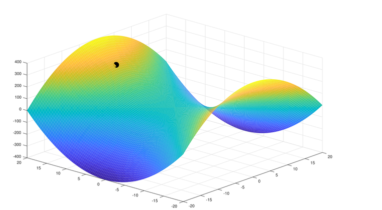
- How should we define an SOSP for the constrained setting?
- $x^* \in \mathcal{C}$ is an approximate (ε, γ)-second-order order stationary point if

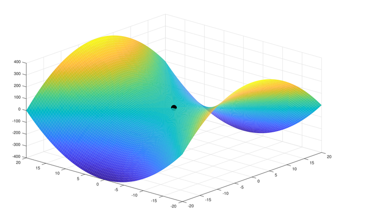
$$\nabla f(x^*)^T(x - x^*) \geq -\varepsilon \quad \text{for all } x \in \mathcal{C}$$

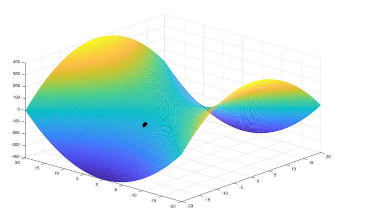
$$(x - x^*)^T \nabla^2 f(x^*)(x - x^*) \geq -\gamma \quad \text{for all } x \in \mathcal{C} \text{ s.t. } \nabla f(x^*)^T(x - x^*) = 0$$

- Second condition should be satisfied only on the subspace that function can be increasing
- Setting $\varepsilon = \gamma = 0$ gives the necessary conditions for a local min
- We propose a framework that finds an (ε, γ)-SOSP in poly-time
  - $\Rightarrow$ If optimizing a quadratic loss over $\mathcal{C}$ up to a constant factor is tractable

Proposed algorithm to find an (ε, γ)-SOSP







- Follow a first-order update to reach an ε-FOSP
  - $\Rightarrow$ The function value decreases at a rate of $\mathcal{O}(\epsilon^{-2})$
- Escape from saddle points by solving a QP which depends objective function curvature information
  - $\Rightarrow$ The function value decreases at a rate of $\mathcal{O}(\gamma^{-3})$
- Once we escape from a saddle point we won't revisit it again
  - $\Rightarrow$ The function value decreases after escaping from saddle
  - $\Rightarrow$ It is guaranteed that the function value never increases

Stage I: First-order update (Finding a critical point)

- Goal: Find x_t s.t. $\Rightarrow \nabla f(x_t)^T(x - x_t) \geq -\varepsilon$ for all $x \in \mathcal{C}$
- Follow Frank-Wolfe until reaching an ε-FOSP

$$x_{t+1} = (1 - \eta)x_t + \eta v_t, \qquad \text{where} \quad v_t = \operatorname*{argmin}_{v \in \mathcal{C}} \{\nabla f(x_t)^T v\}$$

- Follow Projected Gradient Descent until reaching an ε-FOSP

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

$\pi_{\mathcal{C}}(.)$ is the Euclidean projection onto the convex set $\mathcal{C}$

- The function value decreases at least by a factor of $\mathcal{O}(\epsilon^{-2})$

Stage II: Second-order update (Escaping from saddle points)

- Find u_t a ρ-approximate solution of the quadratic program

$$\begin{aligned} \text{Minimize} \quad & q(u) := (u - x_t)^T \nabla^2 f(x_t)(u - x_t) \\ \text{subject to} \quad & u \in \mathcal{C}, \quad \nabla f(x_t)^T(u - x_t) = 0 \end{aligned}$$

- $q(u^*) \leq q(u_t) \leq \rho q(u^*)$ for some $\rho \in (0, 1]$

$$\begin{cases} \text{If } \; q(u_t) < -\rho\gamma & \Rightarrow \; \text{Update } x_{t+1} = (1 - \sigma)x_t + \sigma u_t \\ \text{If } \; q(u_t) \geq -\rho\gamma \; \Rightarrow \; q(u^*) \geq -\gamma & \Rightarrow \; x_t \text{ is an } (\varepsilon, \gamma)\text{-SOSP} \end{cases}$$

- Some classes of convex constraints satisfy this property
  - $\Rightarrow$ Quadratic constraints under some conditions

Theoretical Results

Theorem. *If we set the stepsizes to $\eta = \mathcal{O}(\varepsilon)$ and $\sigma = \mathcal{O}(\rho\gamma)$, the proposed algorithm finds an (ε, γ)-SOSP after at most $\mathcal{O}(\max\{\varepsilon^{-2}, \rho^{-3}\gamma^{-3}\})$ iterations.*

- When can we solve the quadratic subproblem approximately?

Proposition *If $\mathcal{C}$ is defined by a quadratic constraint, then the alg. finds an (ε, γ)-SOSP after $\mathcal{O}(\max\{\tau\varepsilon^{-2}, d^3\gamma^{-3}\})$ arith. operations.*

Proposition *If the convex set $\mathcal{C}$ is defined as a set of m quadratic constraints ($m > 1$), and the objective function Hessian satisfies $\max_{x \in \mathcal{C}} x^T \nabla^2 f(x) x \leq \mathcal{O}(\gamma)$, then the algorithm finds an (ε, γ)-SOSP at most after $\mathcal{O}(\max\{\tau\varepsilon^{-2}, d^3 m^7 \gamma^{-3}\})$ arithmetic operations.*

Proposed Algorithm

```
for t = 1, 2, ...
    Compute v_t = argmin_{v∈C} {∇f(x_t)^T v}
    if ∇f(x_t)^T (v_t − x_t) < −ε
        x_{t+1} = (1 − η)x_t + ηv_t
    else
        Find u_t: a ρ-approximate solution of the QP
        if q(u_t) < −ργ
            x_{t+1} = (1 − σ)x_t + σu_t
        else
            return x_t and stop
```

Stochastic Setting

- What about the stochastic setting?

$$\min_{x \in \mathcal{C}} f(x) = \min_{x \in \mathcal{C}} \mathbb{E}_{\Theta}[F(x, \Theta)]$$

- where Θ is a random variable with probability distribution $\mathcal{P}$
- Replace $\nabla f(x_t)$ and $\nabla^2 f(x_t)$ by their stochastic approximations g_t and H_t

$$g_t = \frac{1}{b_g} \sum_{i=1}^{b_g} \nabla F(x_t, \theta_i), \qquad H_t = \frac{1}{b_H} \sum_{i=1}^{b_H} \nabla^2 F(x_t, \theta_i)$$

- Change some conditions to afford approximation error
  - $\Rightarrow \nabla f(x_t)^T(x - x_t) = 0 \;\Rightarrow\; \nabla g_t^T(x - x_t) \leq r$

Proposed Method for the Stochastic Setting

```
for t = 1, 2, ...
    Compute v_t = argmin_{v∈C} {g_t^T v}
    if g_t^T (v_t − x_t) ≤ −ε/2
        x_{t+1} = (1 − η)x_t + ηv_t
    else
        Find u_t: a ρ-approximate solution of
            min q(u) := (u − x_t)^T H_t (u − x_t)
            s.t. u ∈ C,  g_t^T (u − x_t) ≤ r
        if q(u_t) < −ργ/2
            x_{t+1} = (1 − σ)x_t + σu_t
        else
            return x_t and stop
```

Theoretical Results for the Stochastic Setting

Theorem. *If we set stepsizes to $\eta = \mathcal{O}(\varepsilon)$ and $\sigma = \mathcal{O}(\rho\gamma)$, batch sizes to $b_g = \mathcal{O}(\max\{\rho^{-4}\gamma^{-4}, \varepsilon^{-2}\})$ and $b_H = \mathcal{O}(\rho^{-2}\gamma^{-2})$, and choose $r = \mathcal{O}(\rho^2\gamma^2)$,*

- *$\Rightarrow$ The outcome of Algorithm 2 is an (ε, γ)-SOSP w.h.p.*
- *$\Rightarrow$ Total No. of iterations is at most $\mathcal{O}(\max\{\varepsilon^{-2}, \rho^{-3}\gamma^{-3}\})$ w.h.p.*

Corollary *Algorithm finds an (ε, γ)-SOSP w.h.p. after computing*

- *$\Rightarrow \mathcal{O}(\max\{\varepsilon^{-2}\rho^{-4}\gamma^{-4}, \varepsilon^{-4}, \rho^{-7}\gamma^{-7}\})$ stochastic gradients*
- *$\Rightarrow \mathcal{O}(\max\{\varepsilon^{-2}\rho^{-3}\gamma^{-3}, \rho^{-5}\gamma^{-5}\})$ stochastic Hessians*

Conclusion

- Method for finding an SOSP in constrained settings
  - $\Rightarrow$ Using first-order information to reach an FOSP
  - $\Rightarrow$ Solve a QP up to a constant factor $\rho < 1$ to escape from saddles
- First finite-time complexity analysis for constrained problems
  - $\Rightarrow \mathcal{O}(\max\{\varepsilon^{-2}, \rho^{-3}\gamma^{-3}\})$ iter. $\Rightarrow \mathcal{O}(\max\{\tau\varepsilon^{-2}, d^3m^7\gamma^{-3}\})$ A.O. for QP
  - $\Rightarrow$ Extended our results to the stochastic setting