Non-convex Optimization for Machine Learning

Prateek Jain

Microsoft Research India
Outline

• Optimization for Machine Learning
• Non-convex Optimization
• Convergence to Stationary Points
  • First order stationary points
  • Second order stationary points
• Non-convex Optimization in ML
  • Neural Networks
  • Learning with Structure
    • Alternating Minimization
    • Projected Gradient Descent
Non-convex Optimization for Machine Learning
Prateek Jain and Purushottam Kar
Optimization in ML

Supervised Learning
- Given points \((x_i, y_i)\)
- Prediction function: \(\hat{y}_i = \phi(x_i, w)\)
- Minimize loss: \(\min_w \sum_i \ell(\phi(x_i, w), y_i)\)

Unsupervised Learning
- Given points \((x_1, x_2 \ldots x_N)\)
- Find cluster center or train GANs
- Represent \(\hat{x}_i = \phi(x_i, w)\)
- Minimize loss: \(\min_w \sum_i \ell(\phi(x_i, w), x_i)\)
Optimization Problems

• Unconstrained optimization

\[
\min_{w \in \mathbb{R}^d} f(w)
\]

• Deep networks
• Regression
• Gradient Boosted Decision Trees

• Constrained optimization

\[
\min_{w} f(w) \quad \text{s.t.} \quad w \in C
\]

• Support Vector Machines
• Sparse regression
• Recommendation system
• …
Convex Optimization

\[
\min_{w} f(w) \\
\text{s.t. } w \in C
\]

\[
f \colon \mathbb{R}^d \rightarrow \mathbb{R}
\]

Convex function

\[
f(\lambda w_1 + (1 - \lambda)w_2) \leq \lambda f(w_1) + (1 - \lambda)f(w_2),
\]

\[
0 \leq \lambda \leq 1
\]

\[
C \subseteq \mathbb{R}^d
\]

Convex set

\[
\forall w_1, w_2 \in C, \lambda w_1 + (1 - \lambda)w_2 \in C
\]

\[
0 \leq \lambda \leq 1
\]

Slide credit: Purushottam Kar
Examples

Linear Programming

\[
\min_{x \in \mathbb{R}^d} \ a^\top x \\
\text{s.t. } b_i^\top x \leq c_i
\]

Quadratic Programming

\[
\min_{x \in \mathbb{R}^d} \ \frac{1}{2} x^\top Ax + a^\top x \\
\text{s.t. } b_i^\top x \leq c_i
\]

Semidefinite Programming

\[
\min_{X \succeq 0} \ A^\top X \\
\text{s.t. } B_i^\top X \leq c_i
\]
Convex Optimization

- **Unconstrained optimization**
  \[ \min_{w \in \mathbb{R}^d} f(w) \]
  Optima: just ensure  \( \nabla_w f(w) = 0 \)

- **Constrained optimization**
  \[ \min_w f(w) \text{ s.t. } w \in C \]
  Optima: KKT conditions

In this talk, let's assume \( f \) is \( L \) -smooth \( \Rightarrow \) \( f \) is differentiable

\[
\begin{align*}
f(x) & \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} ||x - y||^2 \\
\text{OR,} & \quad ||\nabla f(x) - \nabla f(y)|| \leq L ||x - y||
\end{align*}
\]
Gradient Descent Methods

• Projected gradient descent method:

  • For $t=1, 2, \ldots$ (until convergence)
    • $w_{t+1} = P_C(w_t - \eta \nabla f(w_t))$

  • $\eta$: step-size
Convergence Proof

\[
f(w_{t+1}) \leq f(w_{t}) + \langle \nabla f(w_{t}), w_{t+1} - w_{t} \rangle + \frac{L}{2} \|w_{t+1} - w_{t}\|^2
\]

\[
f(w_{t+1}) \leq f(w_{t}) - \left(1 - \frac{L\eta}{2}\right)\eta \|\nabla f(w_{t})\|^2 \leq f(w_{t}) - \frac{\eta}{2} \|\nabla f(w_{t})\|^2
\]

\[
f(w_{t+1}) \leq f(w_{*}) + \langle \nabla f(w_{t}), w_{t} - w_{*} \rangle - \frac{1}{2\eta} \|w_{t+1} - w_{t}\|^2
\]

\[
f(w_{t+1}) \leq f(w_{*}) + \frac{1}{2\eta} \left(\|w_{t} - w_{*}\|^2 - \|w_{t+1} - w_{*}\|^2\right)
\]

\[
f(w_{T}) \leq f(w_{*}) + \frac{1}{T \cdot 2\eta} \|w_{0} - w_{*}\|^2 \Rightarrow f(w_{T}) \leq f(w_{*}) + \epsilon
\]

\[
T = O\left(\frac{L \cdot \|w_{0} - w_{*}\|^2}{\epsilon}\right)
\]
Non-convexity?

$$\min_{w \in \mathbb{R}^d} f(w)$$

• Critical points: $$\nabla f(w) = 0$$

• But: $$\nabla f(w) = 0 \nRightarrow$$ Optimality
Local Optima

\[ f(w) \leq f(w'), \forall \|w - w'\| \leq \epsilon \]
First Order Stationary Points

- Defined by: $\nabla f(w) = 0$

- But $\nabla^2 f(w)$ need not be positive semi-definite
First Order Stationary Points

- E.g., $f(w) = 0.5(w_1^2 - w_2^2)$
- $\nabla f(w) = \begin{bmatrix} w_1 \\ -w_2 \end{bmatrix}$
- $\nabla f(0) = 0$
- But, $\nabla^2 f(w) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \Rightarrow$ indefinite

- $f\left(\left[\frac{\epsilon}{2}, \epsilon\right]\right) = -\frac{3}{8} \epsilon^2 \Rightarrow f([0,0])$ is not a local minima
Second Order Stationary Point (SOSP) if:
• $\nabla f(w) = 0$
• $\nabla^2 f(w) \succeq 0$

Does it imply local optimality?
Second Order Stationary Points

- $f(w) = \frac{1}{3}(w_1^3 - 3w_1w_2^2)$
- $\nabla f(w) = \begin{bmatrix} (w_1^2 - w_2^2) \\ -2w_1w_2 \end{bmatrix}$
- $\nabla^2 f(w) = \begin{bmatrix} 2w_1 & -2w_2 \\ -2w_2 & -2w_1 \end{bmatrix}$
- $\nabla f(0) = 0, \nabla^2 f(0) = 0 \Rightarrow 0$ is SOSP
- $f([\epsilon, \epsilon]) = -\frac{2}{3}\epsilon^3 < f(0)$
Stationarity and local optima

- $w$ is local optima implies: $f(w) \leq f(w')$, $\forall ||w - w'|| \leq \epsilon$

- $w$ is FOSP implies:
  $$f(w) \leq f(w') + O(||w - w||^2)$$

- $w$ is SOSP implies:
  $$f(w) \leq f(w') + O(||w - w'||^3)$$

- $w$ is $p$-th order SP implies:
  $$f(w) \leq f(w') + O(||w - w'||^{p+1})$$

- That is, local optima: $p = \infty$
### Computability?

\[ f(w) \leq f(w') + O(||w - w'||^{p+1}) \]

<table>
<thead>
<tr>
<th>Order</th>
<th>Stationary Point</th>
<th>Computability</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Order</td>
<td>✓</td>
<td>Computable</td>
</tr>
<tr>
<td>Second Order</td>
<td>✓</td>
<td>Computable</td>
</tr>
<tr>
<td>Third Order</td>
<td>✓</td>
<td>Computable</td>
</tr>
<tr>
<td>( p \geq 4 )</td>
<td>×</td>
<td>NP-Hard</td>
</tr>
<tr>
<td>Local Optima</td>
<td>×</td>
<td>NP-Hard</td>
</tr>
</tbody>
</table>

Anandkumar and Ge-2016
Does Gradient Descent Work for Local Optimality?

• Yes!

• In fact, with high probability converges to a “local minimizer”
  • If initialized randomly!!!

• But no rates known 😞
  • NP-hard in general!!
  • Big open problem 😊
Finding First Order Stationary Points

- Defined by: $\nabla f(w) = 0$
- But $\nabla^2 f(w)$ need not be positive semi-definite
Gradient Descent Methods

• Gradient descent:

  • For $t=1, 2, \ldots$ (until convergence)
    • $w_{t+1} = w_t - \eta \nabla f(w_t)$

  • $\eta$: step-size

  • Assume:
    $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$
Convergence to FOSP

\[
f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} \|w_{t+1} - w_t\|^2
\]

\[
f(w_{t+1}) \leq f(w_t) - \left(1 - \frac{L\eta}{2}\right) \eta \|\nabla f(w_t)\|^2 \leq f(w_t) - \frac{1}{2L} \|\nabla f(w_t)\|^2
\]

\[
\|\nabla f(w_t)\|^2 \leq f(w_t) - f(w_{t+1})
\]

\[
\frac{1}{2L} \sum_t \|\nabla f(w_t)\|^2 \leq f(w_0) - f(w_*)
\]

\[
\min_t \|\nabla f(w_t)\| \leq \sqrt{\frac{2L}{T} (f(w_0) - f(w_*))} \leq \epsilon
\]

\[
T = O\left(\frac{L \cdot (f(w_0) - f(w_*))}{\epsilon^2}\right)
\]
Accelerated Gradient Descent for FOSP?

- For $t=1, 2,..., T$
  - $w_{t+1}^{md} = (1 - \alpha_t)w_t^{ag} + \alpha_t w_t$
  - $w_{t+1} = w_t - \eta_t \nabla f(w_{t+1}^{md})$
  - $w_{t+1}^{ag} = w_t^{md} - \beta_t \nabla f(w_{t+1}^{md})$

- Convergence? $\min_t ||\nabla f(w_t)|| \leq \epsilon$

- For $T = O\left(\frac{\sqrt{L \cdot (f(w_0) - f(w^*))}}{\epsilon} \right)$

- If convex: $T = O\left(\frac{(L \cdot (f(w_0) - f(w^*))^{1/4}}{\sqrt{\epsilon}} \right)$

Ghadimi and Lan - 2013
Non-convex Optimization: Sum of Functions

• What if the function has more structure?

\[
\min_w f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)
\]

• \( \nabla f(w) = \sum_{i=1}^{n} \nabla f_i(w) \)
• I.e., computing gradient would require \( O(n) \) computation
Does Stochastic Gradient Descent Work?

For $t=1, 2, \ldots$ (until convergence)
  - Sample $i_t \sim Unif[1, n]$
  - $w_{t+1} = w_t - \eta \nabla f_{i_t}(w_t)$

Proof? $E_{i_t}[w_{t+1} - w_t] = \eta \nabla f(w_t)$

\[
\begin{align*}
f(w_{t+1}) &\leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} ||w_{t+1} - w_t||^2 \\
E[f(w_{t+1})] &\leq E[f(w_t)] - \frac{\eta}{2} ||\nabla f(w_t)||^2 + \frac{L}{2} \eta^2 \cdot Var \\
\min_t ||\nabla f(w_t)|| &\leq \left( \frac{L (f(w_0) - f(w_*)) \cdot Var}{T^{\frac{1}{4}}} \right)^{\frac{1}{4}} \leq \epsilon
\end{align*}
\]

\[
T = O \left( \frac{L \cdot Var \cdot (f(w_0) - f(w_*))}{\epsilon^4} \right)
\]
## Summary: Convergence to FOSP

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>No. of Gradient Calls (Non-convex)</th>
<th>No. of Gradient Calls (Convex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD [Folkore; Nesterov]</td>
<td>$O\left(\frac{1}{\epsilon^2}\right)$</td>
<td>$O\left(\frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>AGD [Ghadimi &amp; Lan-2013]</td>
<td>$O\left(\frac{1}{\epsilon}\right)$</td>
<td>$O\left(\frac{1}{\sqrt{\epsilon}}\right)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>No. of Gradient Calls</th>
<th>Convex Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD [Folkore]</td>
<td>$O\left(\frac{n}{\epsilon^2}\right)$</td>
<td>$O\left(\frac{n}{\epsilon}\right)$</td>
</tr>
<tr>
<td>AGD [Ghadimi &amp; Lan’2013]</td>
<td>$O\left(\frac{n}{\epsilon}\right)$</td>
<td>$O\left(\frac{n}{\sqrt{\epsilon}}\right)$</td>
</tr>
<tr>
<td>SGD [Ghadimi &amp; Lan’2013]</td>
<td>$O\left(\frac{1}{\epsilon^4}\right)$</td>
<td>$O\left(\frac{1}{\epsilon^2}\right)$</td>
</tr>
<tr>
<td>SVRG [Reddi et al-2016, Allen-Zhu&amp;Hazan-2016]</td>
<td>$O\left(n + \frac{n^3}{\epsilon^2}\right)$</td>
<td>$O\left(n + \sqrt{n}/\epsilon^2\right)$</td>
</tr>
<tr>
<td>MSVRG [Reddi et al-2016]</td>
<td>$O\left(\min\left(\frac{1}{\epsilon^4}, \frac{n^3}{\epsilon^2}\right)\right)$</td>
<td>$O\left(n + \sqrt{n}/\epsilon^2\right)$</td>
</tr>
</tbody>
</table>

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$
Finding Second Order Stationary Points (SOSP)

Second Order Stationary Point (SOSP) if:
• $\nabla f(w) = 0$
• $\nabla^2 f(w) \succeq 0$

Approximate SOSP:
• $||\nabla f(w)|| \leq \varepsilon$
• $\lambda_{min}(\nabla^2 f(w)) \geq -\sqrt{\rho\varepsilon}$
Cubic Regularization (Nesterov and Polyak-2006)

• For \( t=1, 2, \ldots \) (until convergence)

\[
\begin{align*}
\mathbf{w}_{t+1} &= \arg \min_{\mathbf{w}} f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, \nabla f(\mathbf{w}_t) \rangle + \frac{1}{2} (\mathbf{w} - \mathbf{w}_t)^T \nabla^2 f(\mathbf{w}_t) (\mathbf{w} - \mathbf{w}_t) + \frac{\rho}{6} ||\mathbf{w} - \mathbf{w}_t||^3
\end{align*}
\]

• Assumption: Hessian continuity, i.e., \( ||\nabla^2 f(x) - \nabla^2 f(y)|| \leq \rho ||x - y|| \)

• Convergence to SOSP? \( T = O\left(\frac{1}{\epsilon^{1.5}}\right) \)
  • But requires Hessian computation! (even storage is \( O(d^2) \))
  • Can we find SOSP using only gradients?
Noisy Gradient Descent for SOSP

• For t=1, 2, ... (until convergence)
  • If ( ||∇f(w_t)|| ≥ ε )
    • w_{t+1} = w_t - η∇f(w_t)
  • Else
    • w_{t+1} = w_t + ζ, ζ ~ γ · N(0, I)
    • Update w_{t+1} = w_t − η∇f(w_t) for next r iterations

• Claim: above algorithm converges to SOSP in $O(1/\epsilon^2)$

Ge et al-2015, Jin et al-2017
Proof

For t=1, 2, ... (until convergence)

If ( ||∇f(w_t)|| ≥ ε )

\[ w_{t+1} = w_t - \eta \nabla f(w_t) \]

Else

\[ w_{t+1} = w_t + \zeta, \zeta \sim \gamma \cdot N(0, I) \]

Update \( w_{t+1} = w_t - \eta \nabla f(w_t) \) for next \( r \) iterations

FOSP analysis: convergence in \( O\left(\frac{1}{\epsilon^2}\right) \) iterations

But, \( \nabla^2 f(w_t) \neq 0 \)

• That is, \( \lambda_{min}(\nabla^2 f(w_t)) < -\sqrt{\rho \epsilon} \)

image credit: academo.org
Proof

For $t=1, 2, \ldots$ (until convergence)

If ($||\nabla f(w_t)|| \geq \epsilon$)

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$

Else

$$w_{t+1} = w_t + \zeta, \zeta \sim \gamma \cdot N(0, I)$$

Update $w_{t+1} = w_t - \eta \nabla f(w_t)$ for next $r$ iterations

• Random perturbation with Gradient descent leads to decrease in objective function
Proof?

• Random perturbation with Gradient descent leads to decrease in objective function

• Hessian continuity => function nearly quadratic in small neighborhood

• \( f(w) \approx f(w_t) + (\nabla f(w_t), w - w_t) + (w - w_t)^T \nabla^2 f(w_t)(w - w_t) \)

\[
\begin{align*}
w_{r+1} &= w_{r-1+t} - \eta \nabla^2 f(w_t)(w_{r-1+t} - w_t) \\
\Rightarrow w_{r+1} - w_t &= \left(I - \eta \nabla^2 f(w_t)\right)' (w_{t+1} - w_t)
\end{align*}
\]

For \( t=1, 2, \ldots \) (until convergence)

If \( \|\nabla f(w_t)\| \geq \epsilon \)

\[
w_{t+1} = w_t - \eta \nabla f(w_t)
\]

Else

\[
w_{t+1} = w_t + \zeta, \zeta \sim \gamma \cdot N(0, I)
\]

Update \( w_{t+1} = w_t - \eta \nabla f(w_t) \) for next \( r \) iterations
Proof?

• Random perturbation with Gradient descent leads to decrease in objective function

• Hessian continuity => function nearly quadratic in small neighborhood

\[ f(w) \approx f(w_t) + \langle \nabla f(w_t), w - w_t \rangle + (w - w_t)^T \nabla^2 f(w_t)(w - w_t) \]

\[ w_{r+t} = w_{r-1+t} - \eta \nabla^2 f(w_t)(w_{r-1+t} - w_t) \]

\[ \Rightarrow w_{r+t} - w_t = (I - \eta \nabla^2 f(w_t))^r (w_{t+1} - w_t) \]

• \( w_{r+t} - w_t \) converge to largest eigenvector of \( I - \eta \nabla^2 f(w_t) \)
  • Which is smallest (most negative) eigenvector of \( \nabla^2 f(w_t) \)

• Hence, \( (w_{r+t} - w_t)^T \nabla^2 f(w_t)(w_{r+t} - w_t) \leq -\gamma^2 \sqrt{\rho \epsilon} \)

• \( f(w_{r+t}) \leq f(w_t) - \gamma^2 \sqrt{\rho \epsilon} \)
Proof

For $t=1, 2, \ldots$ (until convergence)

If ($\|\nabla f(w_t)\| \geq \epsilon$)

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$

Else

$$w_{t+1} = w_t + \zeta, \zeta \sim \gamma \cdot N(0, I)$$

Update $w_{t+1} = w_t - \eta \nabla f(w_t)$ for next $r$ iterations

• Entrapment near SOSP

Final result: convergence to SOSP in $O(1/\epsilon^2)$

image credit: academo.org

Ge et al-2015, Jin et al-2017
### Summary: Convergence to SOSP

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>No. of Gradient Calls (Non-convex)</th>
<th>No. of Gradient Calls (Convex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy GD [Jin et al-2017, Ge et al-2015]</td>
<td>$O\left(\frac{1}{\epsilon^2}\right)$</td>
<td>$O\left(\frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>Noisy Accelerated GD [Jin et al-2017]</td>
<td>$O\left(\frac{1}{\epsilon^{1.75}}\right)$</td>
<td>$O\left(\frac{1}{\sqrt{\epsilon}}\right)$</td>
</tr>
<tr>
<td>Cubic Regularization [Nesterov &amp; Polyak-2006]</td>
<td>$O\left(\frac{1}{\epsilon^{1.5}}\right)$</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Algorithm

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$
Convergence to Global Optima?

- FOSP/SOSP methods can’t even guarantee local convergence

- Can we guarantee global optimality for some “nicer” non-convex problems?
  - Yes!!!
  - Use statistics 😊
Can Statistics Help: Realizable models!

- Data points: \((x_i, y_i) \sim D\)
- \(D\): nice distribution
- \(E[y_i] = \phi(x_i, w_*)\)

\[ \hat{w} = \arg \min_{w} \sum_i \text{loss}(y_i, \phi(x_i, w)) \]

- That is, \(w_*\) is the optimal solution!
  - Parameter learning
Learning Neural Networks: Provably

• $y_i = 1 \cdot \sigma(W_* x_i)$
• $x_i \sim N(0, I)$

$$\min_W \sum_i \left( y_i - 1 \cdot \sigma(W x_i) \right)^2$$

• Does gradient descent converge to global optima: $W_*$?
  • NO!!!
  • The objective function has poor local minima [Shamir et al-2017, Lee et al-2017]
Learning Neural Networks: Provably

• But, no local minima within constant distance of $W_*$
• If,

$$||W_0 - W_*|| \leq c$$

Then, Gradient Descent ($W_{t+1} = W_t - \eta \nabla f(W_t)$) converges to $W_*$

No. of iterations: $\log \frac{1}{\epsilon}$

Can we get rid of initialization condition? Yes but by changing the network [Liang-Lee-Srikant’2018]
Learning with Structure

• \( y_i = \phi(x_i, w_*) \), \( x_i \sim D \in \mathbb{R}^d \), \( 1 \leq i \leq n \)

• But no. of samples are limited!
  • For example, \( \text{if } n \leq d ? \)

• Can we still recover \( w_* \)? In general, no!
  • But, what if \( w_* \) has some structure?
Sparse Linear Regression

- But: \( n \ll d \)
- \( w: s - \text{sparse} (s \text{ non-zeros}) \)
  - Information theoretically: \( n = s \log d \) samples should suffice
Learning with structure

\[
\min_{w} f(w) \\
\text{s.t. } w \in C
\]

• Linear classification/regression
  • \( C = \{w, \|w\|_0 \leq s\} \)
  • \( s \ll d \)

• Matrix completion
  • \( C = \{W, \text{rank}(W) \leq r\} \)
  • \( r \ll (d_1, d_2) \)
Other Examples

• Low-rank Tensor completion
  • $C = \{ W, \text{ tensor} - \text{rank}(W) \leq r \}$
  • $r \ll (d_1, d_2, d_3)$

• Robust PCA
  • $C = \{ W, W = L + S, \text{rank}(L) \leq r, ||S||_0 \leq s \}$
  • $r \ll (d_1, d_2), S \ll d_1 \times d_2$
Non-convex Structures

• Linear classification/regression
  • $\mathcal{C} = \{w, \|w\|_0 \leq s\}$
  • $s \ll d$

• Matrix completion
  • $\mathcal{C} = \{W, \text{rank}(W) \leq r\}$
  • $r \ll (d_1, d_2)$

• NP-Hard
  • $\|w\|_0$: Non-convex

• NP-Hard
  • $\text{rank}(W)$: Non-convex
Non-convex Structures

• Low-rank Tensor completion
  • \( C = \{ W, \text{ tensor } \text{–} \text{rank} (W) \leq r \} \)
  • \( r \ll (d_1, d_2, d_3) \)

• Robust PCA
  • \( C = \{ W, W = L + S, \text{rank} (L) \leq r, \|S\|_0 \leq s \} \)
  • \( r \ll (d_1, d_2), S \ll d_1 \times d_2 \)

  • Indeterminate
  • \( \text{tenorsrank}(W) \): Non-convex

  • NP-Hard
  • \( \text{rank}(W), \|S\|_0 \): Non-convex
Technique: Projected Gradient Descent

\[
\begin{align*}
\min_{w} & \quad f(w) \\
\text{s.t.} & \quad w \in C
\end{align*}
\]

- \( w_{t+1} = w_t - \nabla_w f(w_t) \)
- \( w_{t+1} = P_C(w_{t+1}) \)

\[
\begin{align*}
\min_{w} & \quad ||w - w_{t+1}||^2 \\
\text{s.t.} & \quad w \in C
\end{align*}
\]
Results for Several Problems

• Sparse regression [Jain et al.’14, Garg and Khandekar’09]
  • Sparsity

• Robust Regression [Bhatia et al.’15]
  • Sparsity+output sparsity

• Vector-value Regression [Jain & Tewari’15]
  • Sparsity+positive definite matrix

• Dictionary Learning [Agarwal et al.’14]
  • Matrix Factorization + Sparsity

• Phase Sensing [Netrapalli et al.’13]
  • System of Quadratic Equations
Results Contd...

• Low-rank Matrix Regression [Jain et al.’10, Jain et al.’13]
  • Low-rank structure

• Low-rank Matrix Completion [Jain & Netrapalli’15, Jain et al.’13]
  • Low-rank structure

• Robust PCA [Netrapalli et al.’14]
  • Low-rank ∩ Sparse Matrices

• Tensor Completion [Jain and Oh’14]
  • Low-tensor rank

• Low-rank matrix approximation [Bhojanapalli et al.’15]
  • Low-rank structure
Sparse Linear Regression

\[
y = Xw
\]

- But: \( n \ll d \)
- \( w \): \( s \) -sparse (\( s \) non-zeros)
Sparse Linear Regression

\[
\min_w \|y - Xw\|^2 \\
\text{s.t. } \|w\|_0 \leq s
\]

• \(\|y - Xw\|^2 = \sum_i (y_i - \langle x_i, w \rangle)^2\)
• \(\|w\|_0\): number of non-zeros

• NP-hard problem in general 😞
  • \(L_0\): non-convex function
Technique: Projected Gradient Descent

\[
\min_{w} f(w) = \|y - Xw\|^2
\]
\[
s.t. \quad \|w\|_0 \leq s
\]
\[
\cdot w_{t+1} = w_t - \nabla_w f(w_t)
\]
\[
\cdot w_{t+1} = P_s(w_{t+1})
\]
Statistical Guarantees

\[ y_i = \langle x_i, w^* \rangle + \eta_i \]

- \( x_i \sim N(0, \Sigma) \)
- \( \eta_i \sim N(0, \zeta^2) \)
- \( w^* : s \)-sparse

\[ || \hat{w} - w^* || \leq \frac{\zeta \kappa^3 \sqrt{s \log d}}{\sqrt{n}} \]

- \( \kappa = \lambda_1(\Sigma) / \lambda_d(\Sigma) \)

[Jain, Tewari, Kar’2014]
Low-rank Matrix Completion

\[
\min_W \sum_{(i,j) \in \Omega} (W_{ij} - M_{ij})^2 \\
\text{s. t. } \text{rank}(W) \leq r
\]

\(\Omega\): set of known entries

• Special case of low-rank matrix regression
• However, assumptions required by the regression analysis not satisfied
Technique: Projected Gradient Descent

- $W_0 = 0$
- For $t=0:T-1$
  \[ W_{t+1} = P_r(W_t - \eta \nabla f(W_t)) \]
- $P_k(Z)$: projection onto set of rank-$r$ projection
- Singular Value Projection
- Pros:
  - Fast (always, rank-$r$ SVD)
  - Matrix completion: $O(d \cdot r^3)$!
- Cons: In general, might not even converge
- Our Result: Convergence under “certain” assumptions

[Jain, Tewari, Kar’2014], [Netrapalli, Jain’2014], [Jain, Meka, Dhillon’2009]
Guarantees

• Projected Gradient Descent:
  • $W_{t+1} = Pr\left(W_t - \eta \nabla_W f(W_t)\right)$, $\forall t$

• Show $\epsilon$-approximate recovery in $\log \frac{1}{\epsilon}$ iterations

• Assuming:
  • $M$: incoherent
  • $\Omega$: uniformly sampled
  • $|\Omega| \geq n \cdot r^5 \cdot \log^3 n$

• First near linear time algorithm for **exact** Matrix Completion with finite samples

[J., Netrapalli’2015]
General Result for Any Function

- $f: \mathbb{R}^d \to \mathbb{R}$
- $f$: satisfies RSC/RSS, i.e.,

$$\alpha \cdot I_{d \times d} \leq H(w) \leq L \cdot I_{d \times d}, \quad \text{if, } w \in C$$

- PGD guarantee:

$$f(w_T) \leq f(w^*) + \epsilon$$

After $T = O(\log \left( \frac{f(w_0)}{\epsilon} \right))$ steps

- If $\frac{L}{\alpha} \leq 1.5$

[J., Tewari, Kar’2014]
Learning with Latent Variables

\[ \min_{w,z} f(w, z) \]

• Typically, \( z \) are latent variables
• E.g., clustering: \( w \): means of clusters, \( z \): cluster index

• \( f \): non-convex
  • NP-hard to solve in general
Alternating Minimization

\[ z_{t+1} = \arg \min_z f(w_t, z) \]
\[ w_{t+1} = \arg \min_w f(w, z_{t+1}) \]

• For example, if \( f(w_t, z) \) is convex and \( f(w, z_t) \) is convex
• Does that imply \( f(w, z) \) is convex?
  • No!!!
  • \( f(w, z) = w \cdot z \)
    • Linear in both \( w, z \) individually
• So can Alt. Min. converge to global optima?
Low-rank Matrix Completion

\[ \min_W \sum_{(i,j) \in \Omega} (W_{ij} - M_{ij})^2 \]
\[ \text{s.t. } \text{rank}(W) \leq r \]
\( \Omega \): set of known entries

- Special case of low-rank matrix regression
- However, assumptions required by the regression analysis not satisfied
Matrix Completion: Alternating Minimization

\[ \min_{V} \| y - X \cdot (U^t V^T) \|_2^2 \]

\[ V^{t+1} = \min_{V} \| y - X \cdot (U^t V^T) \|_2^2 \]

\[ U^{t+1} = \min_{U} \| y - X \cdot (U(V^{t+1})^T) \|_2^2 \]
Results: Alternating Minimization

• Provable global convergence [J., Netrapalli, Sanghavi’13]
• Rate of convergence: geometric
  \[ \|W_T - W^*\| \leq 2^{-T} \]
• Assumptions:
  • Matrix regression: RIP
  • Matrix completion: uniform sampling and no. samples \(|\Omega| \geq O(dk^6)\)

[Jain, Netrapalli, Sanghavi’13]
General Results

\[
\min_{w,z} f(w, z)
\]

• Alternating minimization: optimal?
• If:
  • Joint Restricted Strong Convexity (Strong convexity close to the optimal)
  • Restricted Smoothness (smoothness near optimal)
  • Cross-product bound:
    \[
    |\langle w - w_*, \nabla_w f(w, z) - \nabla_w f(w, z_*) \rangle - \langle z - z_*, \nabla_z f(w, z) - \nabla_z f(w_*, z) \rangle| \\
    \leq O(|w - w_*|^2 + |z - z_*|^2)
    \]

Ha and Barber-2017, Jain and Kar-2018
Summary I

Non-convex Optimization: two approaches

1. General non-convex functions
   a. First Order Stationary Point
   b. Second Order Stationary Point

2. Statistical non-convex functions: learning with structure
   a. Projected Gradient Descent (RSC/RSS)
   b. Alternating minimization/EM algorithms (RSC/RSS)
Summary II

• First Order Stationary Point: \( f(w) \leq f(w') + \|w - w'\|^2 \)
  • Tools: gradient descent, acceleration, stochastic gd, variance reduction
  • Key quantity: iteration complexity
  • Several questions: for example, can we do better? Especially in finite sum setting

• Second order stationary point: \( f(w) \leq f(w') + \|w - w'\|^3 \)
  • Tools: noise+gd, noise+acceleration, noise+sgd, noise+variance reduction
  • Several questions: better rates? Can we remove Lipschitz condition on Hessian?
Summary III

• Projected Gradient Descent
  • Works under statistical conditions like RSC/RSS
  • Still several open questions for most problems
  • E.g., tight guarantees support recovery for sparse linear regression?

• Alternating minimization
  • Works under some assumptions on $f$
  • What is the weakest condition on $f$ for Alt. Min. to work?