Large-Scale Optimization

L3. STOCHASTIC GRADIENT DESCENT - 1
Unconstrained Minimization

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

Setting:

• \(f\) is (continuously) differentiable \(\rightarrow\) gradient descent
• \(f\) is not differentiable \(\rightarrow\) subgradient descent
• In both, \(f\) is convex
  - When \(f\) is nonconvex, most of our argument holds, but extra care is needed to distinguish local maxima, local minima, and saddle points
Gradient Descent
Gradient Descent

One of the simplest & most popular optimization algorithm

For $k=1,2,...$

- Choose a search direction
- Choose a step length
- Take a step

$s_k = -\nabla f(x_k)$

$\alpha_k$

$x_{k+1} = x_k + \alpha_k s_k$
Linear Regression

Data: \( \{(x^j, y^j)\}_{j=1}^m, \ x^j \in \mathbb{R}^n, \ y^j \in \mathbb{R} \)

\[
f(w) = \sum_{j=1}^{m} (y^j - w^T x^j)^2 = \|y - Xw\|^2
\]
Linear Regression

\[
f(w) = \sum_{j=1}^{m} (y^j - w^T x^j)^2 = \| y - Xw \|^2
\]

\[
\nabla f(w) = -2 \sum_{j=1}^{m} x^j (y^j - w^T x^j)
\]

\[
= -2X^T(y - Xw)
\]

Check: dimension
When \( f \) is convex, \( x^* \) is a minimizer of

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

if and only if

\[
\nabla f(x^*) = 0
\]

So we can check

\[
\|\nabla f(x_k)\| < \epsilon
\]
Considerations for Large n or m

Scaling: is \( f(w) = \|y - Xw\|^2 \) a right choice for large m?

Stopping criterion: which norm is better?

Finding the correct solution: when the problem become easier / harder?

- Noise
- Number of samples
Steepest Descent View of GD

We want: \( f(x_k + s_k) < f(x_k) \)

\[
f(x_k + s_k) \approx f(x_k) + \nabla f(x_k)^T s_k
\]

\[
s_k = \arg\min_{s \in \mathbb{R}^n} \nabla f(x_k)^T s
\]

\[
= -\nabla f(x_k)
\]

\[
\nabla f(x_k)^T s_k = -\|\nabla f(x_k)\|^2 < 0
\]

if \( \nabla f(x_k) \neq 0 \)

\[
x_{k+1} = x_k + \alpha_k s_k
\]
Another View of GD

\[ f(x_{k+1}) \approx f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2\alpha_k} \|x_{k+1} - x_k\|^2 \]

Minimize RHS for \( x_{k+1} \): optimality cond gives

\[ \nabla f(x_k) + (x_{k+1} - x_k)/\alpha_k = 0 \]

That is,

\[ x_{k+1} = x_k - \alpha_k \nabla f(x_k) \]
Convergence Rate of GD

Convergence of GD depends on

• How to choose stepsizes
  - Exact linesearch
  - Inexact linesearch
  - Fixed stepsize
  - Decreasing stepsize

• Properties of the objective function
  - Lipschitz continuity of gradients
  - Strong convexity of obj. function
Fixed Stepsize

Take $\alpha_k \equiv c$ for a constant $c$ for all iterations

- $c$ too large
- $c$ too small
- Not easy to guess the “right” size of steps

Slow!
Exact Linesearch

Given $x_k$ and $s_k$, choose $\alpha_k$ s.t.

$$\alpha_k = \arg \min_{\alpha > 0} f(x_k + \alpha s_k)$$

In general, solving this problem is not trivial

If $f$ is quadratic, we have an analytic solution:

$$f(x + \alpha s) = \frac{1}{2}(x + \alpha s)^T H(x + \alpha s) + c^T (x + \alpha s)$$

$$\frac{\partial f(x + \alpha s)}{\partial \alpha} = s^T H(x + \alpha s) + c^T s = 0$$

$$\alpha_k = \frac{-\nabla f(x_k)^T s_k}{s_k^T H s_k}$$
Inexact Linesearch

Find an approximate solution to the line search problem

• Wolfe Linesearch
  - Find a stepsize satisfying “sufficient decrease” & “curvature” conditions

• Backtracking (Armijo) Linesearch
  - Simpler version using only the “sufficient decrease” condition
  - In practice, it performs as good as more complicated linesearch methods
Backtracking Linesearch

Find alpha satisfying “Armijo condition”:

\[
f(x_k + \alpha s_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^T s_k, \quad c_1 \in (0, 1)
\]
Backtracking LS Algorithm

Given alpha_0, beta in (0, 1)

- alpha = alpha_0
- For k=1,2,…, maxiter,
  - Check if alpha satisfies Armijo condition. If yes, return it.
  - Otherwise, \( \alpha = \beta \alpha \)
What is the cost of backtracking LS?

- For $k=1,2,\ldots$, maxiter,
  - Check if alpha satisfies Armijo condition. If yes, return it.
    \[ f(x_k + \alpha s_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^T s_k, \quad c_1 \in (0, 1) \]

- Compute once: $f(x_k)$, $\nabla f(x_k)$, $\nabla f(x_k)^T s_k$
- Compute multiple times: $f(x_k + \alpha s_k)$
  - For different values of alpha

- Practicality of LS depends on how costly function evaluations are
Two Important Properties of the Objective Function

1. Lipschitz continuity of the obj. gradient
2. Strong convexity of the obj. function

• These two properties simply the analysis of optimization algorithms, and therefore very popular

• For many statistical data analysis problems, these are computable in theory. In practice?

• In a sense, these two properties are two sides of the same coin, as we’ll see later
Lipschitz Continuity

Let \( f : \mathbb{R}^n \to \mathbb{R} \) is differentiable
\( \nabla f \) is Lipschitz continuous if

\[
\| \nabla f(y) - \nabla f(x) \| \leq L \| y - x \|, \quad \forall x, y \in \mathbb{R}^n
\]

for some \( L > 0 \) (called the Lipschitz constant)

It is equivalent to the following conditions:

\[
f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \| y - x \|^2, \quad \forall x, y \in \mathbb{R}^n
\]

\[
\langle \nabla f(y) - \nabla f(x), y - x \rangle \leq L \| y - x \|^2
\]
**Strong Convexity**

Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is differentiable

\( f \) is strongly convex if

\[
f((1 - \lambda)y + \lambda x) \leq (1 - \lambda)f(y) + \lambda f(x) - \frac{\alpha}{2} \lambda(1 - \lambda)\|y - x\|^2
\]

\( \forall x, y \in \mathbb{R}^n, \forall \lambda \in [0, 1], \text{ for some } \alpha > 0 \)

Aka \( f \) is \( \alpha \)-strongly convex

Equivalent to the following conditions:

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n
\]

\[
\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \alpha \|y - x\|^2
\]
Again,

\(\nabla f\) is Lipschitz continuous with a constant \(L\) iff

\[
f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n
\]

\[
\langle \nabla f(y) - \nabla f(x), y - x \rangle \leq L \|y - x\|^2
\]

\(f\) is \(\alpha\) - strongly convex iff

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n
\]

\[
\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \alpha \|y - x\|^2
\]
What do they mean?

Lipschitz Continuity

Growth of $\nabla f$ is bounded above by a quadratic function

Strong Convexity

Growth of $f$ is bounded below by a quadratic function
How realistic are they?

Lipschitz continuity of $\nabla f$

- If $f$ is twice differentiable, it implies $\nabla^2 f(x) \preceq LI_{n \times n}$

Strong convexity of $f$

- If $f$ is twice differentiable, it implies $\nabla^2 f(x) \succeq \alpha I_{n \times n}$

Ex. linear regression

- $f(w) = \frac{1}{2} \| y - X w \|^2 : \nabla f(w) = -X^T(y - Xw)$, $\nabla^2 f(x) = X^T X$
- $L = \sigma^2_{\max}(X) = \| X \|_2^2$, which is $>0$ unless $X=0$
- $\alpha = \sigma^2_{\min}(X) > 0$?
  - If $m > n$ (tall) and columns of $X$ are linearly independent, yes
  - If $m < n$ (fat), $\sigma_{\min}(X) = 0$
Convergence Rate of GD

We assume that
- $\nabla f$ is Lipschitz continuous with a constant $L > 0$
- $f$ is convex

$$f(x_k) - f(x^*)$$

<table>
<thead>
<tr>
<th></th>
<th>Fixed Stepsize</th>
<th>Backtracking LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$ Convex</td>
<td>$c \leq 1/L$</td>
<td>Sublinear $O(1/k)$</td>
</tr>
<tr>
<td></td>
<td>Sublinear $O(1/k)$</td>
<td>Sublinear $O(1/k)$</td>
</tr>
<tr>
<td>$f$ Strongly Convex</td>
<td>$c \leq 2/(\alpha + L)$</td>
<td>Linear $O(\gamma^k)$, $\gamma \in (0, 1)$</td>
</tr>
<tr>
<td></td>
<td>Linear $O(\gamma^k)$, $\gamma \in (0, 1)$</td>
<td></td>
</tr>
</tbody>
</table>

$\gamma$ depends on the condition number of the Hessian, i.e. $L/\alpha = \lambda_{\text{max}}(\nabla^2 f(x))/\lambda_{\text{min}}(\nabla^2 f(x))$
Ex. Finding $\epsilon = 10^{-12}$ soln

Gradient descent with fixed stepsizes

Convex
Strongly Convex