Today:
- Projects
- Optimality Conditions for smooth convex optim
- Gradient Descent
- Optimality Conditions for non-smooth convex optim
- Subdifferentials
- Subgradient Descent
Smooth convex optimization

If, our convex is differentiable: $\nabla f(x)$ exists at every $x$ in our domain of optimization.

Unconstrained Optim

\[ \min_{x \in \mathbb{R}^d} f(x) \text{ or just } \min_x f(x) \]

Optimality conditions

Unconstrained: $\nabla f(x^*) = 0$

Sufficient & necessary

Constrained Optim

\[ \min_{x \in C} f(x) \]

where $C$ is convex.
Sufficiency \((\nabla f(x) = 0 \Rightarrow x \text{ is a minimizer})\)

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

Recall \(f\) is convex and differentiable \(\Rightarrow\)

for all \(y\)

\[ f(y) \geq f(x) + \nabla f(x)^T (y - x) \]

so \(\nabla f(x) = 0 \Rightarrow\)

\[ f(y) \geq f(x) \]

so \(x\) is a minimizer.
Constrained Optimization

\[ x^* = \arg \min_{x \in C} f(x) \]

\[ \iff \]

\[ \langle \nabla f(x^*), y - x^* \rangle \geq 0 \]

for all \( y \in C \)

Sufficiency:

Assume \( \langle \nabla f(x), y - x \rangle \geq 0 \) for all \( y \in C \)
then
\[ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \geq f(x) \]
so \( x^* \) is a minimizer
Aside \textit{inner-products} & \textit{Cauchy-Schwarz ineq.}

\[
\langle x, y \rangle \leq \|x\|_2 \|y\|_2
\]

\[
x^T x = \|x\|_2^2
\]

\textit{Cauchy-Schwarz ineq.}  \hspace{1cm} \textit{(C-S ineq.)}

higher-dim analog of the fact that

\[
|ab| \leq |a||b|
\]

\[
\Rightarrow \left| \frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2} \right| \leq 1
\]

\[
\Rightarrow \text{can interpret } \frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2} \text{ as } \cos \angle x \sim y
\]

and we have

\[
\langle x, y \rangle = \left( \frac{\cos \angle x \sim y}{\|x\|_2 \|y\|_2} \right) \|x\|_2 \|y\|_2
\]
Note the optimality conditions give us measures of optimality

\[ \text{Unconst } \quad \| \nabla f(x) \|_a \to 0 \quad \text{as } x \to x^* \quad (\star) \]

\[ \text{Const } \quad \text{dist}(\nabla f(k), S) \to 0 \quad \text{(hard-core)} \]
Optimization Alg for smooth convex optim: Gradient Descent

The basic idea is to note that

\[ f(y) \approx f(x) + \nabla f(x)^T (y-x) \]

so if we want to decrease \( f \) by moving to a point \( y = x + \Delta \), we have

\[ f(y) \approx f(x) + \nabla f(x)^T \Delta \]

taking \( \Delta = -\alpha \nabla f(x) \), we have

\[ f(y) \approx f(x) - \alpha \| \nabla f(x) \|_2^2 \leq f(x) \]

Caveat: 1) \( f \) is nice
2) \( \alpha \) is small enough that our linear approx is accurate
Gradient Descent Alg

Input: \( f - \) convex function on \( \mathbb{R}^d \)

\( \nabla f - \) gradient

\( T - \) number of iterations

\( x_0 - \) starting guess

\( \alpha_1, \ldots, \alpha_T - \) step sizes for each iteration

Output: \( x_1, \ldots, x_T \in \mathbb{R}^d \) - estimates of \( x^* \)

for \( t = 1 \) to \( T \)

\[ x_{t+1} = x_t - \alpha_{t+1} \nabla f(x_t) \]
For now we want to show G-D converges under some reasonable assumptions on $f$, that tell us
- how to choose our stepsizes
- how many steps we need to take to converge in the sense that

$$f(x_T) \leq f(x_*) + \epsilon$$
A reasonable assumption is that our function's gradient is at least Lipschitz-continuous:

$$\| \nabla f(x) - \nabla f(y) \|_2 \leq \beta \| x - y \|_2$$

We call the gradient $\beta$-Lipschitz and we call the function $\beta$ smooth.

**Facts**

1) \( \| \nabla f(x) \|_2 \leq \beta \| x - x^* \|_2 \) \hspace{1cm} \text{\textit{\textbf{take} } y = x^* \hspace{1cm} \text{\textit{\textbf{and note that} } \nabla f(y) = \nabla f(x^*)}} \)

2) This implies a concrete bound on how good the approximation

$$ f(y) \approx f(x) + \nabla f(x)^T (y - x) $$

is
\[ \left| f(g) - f(x) - \langle \nabla f(x), y-x \rangle \right| \leq \frac{\beta}{2} \| x - y \|_2^2 \]

"f is $\beta$-smooth"
Now assume $f$ is $\beta$-smooth. Take $\alpha_t = \frac{1}{\beta}$

Then we see

$$f(x_t - \frac{1}{\beta} \nabla f(x_t))$$

$$- \left[ f(x_t) + \langle \nabla f(x_t), -\frac{1}{\beta} \nabla f(x_t) \rangle \right]$$

$$\leq \frac{\beta}{2} \left\| -\frac{1}{\beta} \nabla f(x_t) \right\|^2 = \frac{\beta}{2} \beta^2 \left\| \nabla f(x_t) \right\|^2$$

$$= \frac{1}{2\beta} \left\| \nabla f(x_t) \right\|^2_a$$
\[ f(x_t - \frac{1}{\beta} \nabla f(x_t)) - \left[ f(x_t) + \langle \nabla f(x_t), -\frac{1}{\beta} \nabla f(x_t) \rangle \right] \leq \frac{1}{2\beta} \| \nabla f(x_t) \|_2^2 \]

Note \[ \langle \nabla f(x_t), -\frac{1}{\beta} \nabla f(x_t) \rangle = -\frac{1}{\beta} \| \nabla f(x_t) \|_2^2 \]

\[ \Rightarrow \]

\[ f(x_t - \frac{1}{\beta} \nabla f(x_t)) - f(x_t) \leq \frac{1}{2\beta} \| \nabla f(x_t) \|_2^2 \]

So taking a step of size \( \frac{1}{\beta} \) in the negative gradient direction decreases \( f \).
Note: we didn’t make use of the fact that the function is convex.


Fact: If $f$ is convex & $\beta$-smooth then gradient descent with $x_t = \frac{t}{\beta}$ guarantees that

$$f(x_t) - f(x^*) \leq \frac{2\beta \|x_0 - x^*\|^2}{t}$$

Rate of convergence is $O\left(\frac{1}{t}\right)$.
In general we don't know $\|x_0 - x^*\|_2^2$ so we usually express this result as

$$f(x_T) - f(x^*) \leq \varepsilon$$

which is achieved when

$$T = O\left(\frac{\bar{F}}{\varepsilon^2}\right)$$

iterations.
What about constrained optimization?

Projected Gradient Descent

Inputs: \( f \), \( \nabla f \), \( T \), \( x_0 \), \( x_1, \ldots, x_T \)

\( P_C \) - projection operator onto \( C \)

Output: \( x_1, \ldots, x_T \)

Alg:

for \( t \) in 1 to \( T \):

\[
x_{t+1} = P_C(x_t - \alpha_t \nabla f(x_t))
\]

\( x^* = \arg\min_{x \in C} f(x) \)

\( P_C(x) = \arg\min_{z \in C} \|x - z\|^2 \)
Non-smooth convex optimization

$f$ is convex but not differentiable

Ex:
$$F(\omega) = \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \omega^T x_i)_+ + \lambda \|\omega\|_1$$

In this case, we have no gradient, so can't move in the negative gradient direction to minimize $F$.

Soln: move in the negative dir of a subgradient
Recall for a smooth convex function:
\[ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \]

If \( f \) is convex but not smooth, it may have multiple tangent lines:

Consider \( x^+ \)

In general, we define the subdifferential of a function \( f \) at \( x \):

\[ \text{multiple tangent lines} \]
\[ \partial f(x) = \{ u \mid \text{for all } y, \]
\[ f(y) \geq f(x) + \langle u, y - x \rangle \}\]

This is a set

**Fact** For convex functions, the subdifferential is nonempty, and if \( f \) is differentiable,

\[ \partial f(x) = \{ \nabla f(x) \} \]

We call the vectors in \( \partial f(x) \) subgradients
Example: subdifferential of positive part $f(x) = x_+$

Clearly when $x < 0$, $\partial f(x) = \{ \frac{x}{2} \}$

$x > 0$, $\partial f(x) = \frac{x}{2} I_x$

when $x = 0$

$\partial f(0) = \{ y : f(y) \geq f(0) + y g \}$

for all $y$.

$= \{ \frac{x}{2} : y_+ \geq y g \}$

for all $y$. 

Convince yourself that $\partial f(0) = [0, 1]$

$$
\partial f(x) = \begin{cases} 
1 & \text{if } x > 0 \\
[0, 1] & \text{if } x = 0 \\
0 & \text{if } x < 0 
\end{cases}
$$
Optimality Conditions

Unconstrained
\[ x^* = \arg\min_x f(x) \]
\[ \iff 0 \in \partial f(x^*) \]

(Necessary & Sufficient)

Constrained
\[ x^* = \arg\min_{x \in C} f(x) \]
\[ \iff \text{there is a subgradient } g \text{ in the subdifferential at } x^*, \]
\[ g \in \partial f(x^*), \text{ for which} \]
\[ \langle g, y - x^* \rangle \geq 0 \]
\[ \text{for all } y \in C \]
Ex \( f(x) = |x| \)

\[
\forall f(x) = \begin{cases} 
1 & x > 0 \\
-1 & x < 0 \\
0 & x = 0
\end{cases}
\]