Gradient Descent

Convex fn. means defn 1st order, \( f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle \)
and relu: \( H(f) > 0 \).

Unconstrained min / constrained min
\[ K = \text{Convex set} \quad \forall x \in K \forall y, z \in K \forall \lambda \in [0,1] \]

(1) If differentiable, convex, Lipschitz \( \forall y \) in \( \| f(x) - f(y) \| \leq G \| x-y \| \).
(2) \( \| \nabla f(x) \| \leq G + \| x \| \in K \)

Convex ext to conining set is the non-differentiable case via subgradient.

Unconstrained: \( f(x^*) \) is minimizer if \( \nabla f(x^*) = 0 \). Local = global.
Constrained: \( f(x^*) \) is minimizer if \( \forall y, \langle \nabla f(x^*), y-x^* \rangle \geq 0 \). "Gradient ascent dir as all optima."

Pf: \( f(y) \geq f(x^*) + \langle \nabla f(x^*), y-x^* \rangle \geq f(x^*) \).

So if closed form maybe able to find minimizer by differentatin. But not if unconstrained, or no closed from.

Claim: if \# steps \( T \geq \left( \frac{GD}{\varepsilon^2} \right)^2 \) then
\[ f(x^*) - f(x_T) \leq \varepsilon \]

Here \( x_T = x_T^* = \frac{1}{T} \sum x_t \).

Pf:

Let's look at how
\[ f(x_t) - f(x^*) \] behaves over time.

Define \( \Phi_t = \frac{1}{2\eta} \| x_t - x^* \|^2 \).

Suppose \( f(x_t) - f(x^*) + \Delta \Phi_t \leq \text{blah} \). \( \forall t \)

\[ \sum_t \left( f(x_t) - f(x^*) \right) \leq T \text{, blah + } \sum_t \left( \Phi_{\text{init}} - \Phi_{\text{final}} \right) \]

\[ = f(x^*) - f(x^*) \leq \frac{1}{T} \sum_t \left( f(x_t) - f(x^*) \right) \leq \text{blah + } \frac{\Phi_{\text{init}}}{T} \]

\[ \frac{1}{T} \sum_t \]

algebra reduces \( \leq \varepsilon \).
\[
f(x_k) - f(x^*) + \Delta \Delta_k \rightarrow \frac{1}{2\eta} \left[ \| x_k^+ - x^* \|^2 - \| x_k - x^* \|^2 \right]
\]

\[
= \frac{1}{2\eta} \left[ \| x_k^+ - x_k \|^2 + 2\langle x_k^+ - x_k, x_k - x^* \rangle \right]
\]

by def of GD,

\[
= \frac{1}{2\eta} \left[ \| \nabla f(x_k) \|^2 + 2\langle \nabla f(x_k), x_k - x^* \rangle \right]
\]

\[
= \frac{1}{2\eta} \nabla^2 - \langle \nabla f(x_k), x_k - x^* \rangle.
\]

\[
= f(x_k) - f(x_k^*) - \langle \nabla f(x_k), x_k - x^* \rangle + \frac{\eta}{2} \nabla^2.
\]

\[
\leq \frac{\eta}{2} \nabla^2 \leq \frac{\eta}{2} G^2.
\]

\[
f(x_k) \leq f(x_k^*) - \frac{1}{2} \| x_k - x^* \|^2 - \frac{1}{2\eta} \| \nabla^2 \|^2.
\]

set \( \eta = \frac{D}{c^2} \) to get

\[
DG \frac{1}{2c^2} + DG \frac{1}{2c^2} = DG \leq \epsilon \quad \text{if} \quad T \geq \left( \frac{D^2}{c^2} \right)^2.
\]

So, if we want to be within \( \epsilon \) of \( x^* \), need \( \frac{1}{c^2} \) times. Slow. For 1 move to 1 info, need to run for 4 times more.

But I think example that shows basic GD (not good for further refinement) does \( \frac{1}{c^2} \).

However: constrained opt: take step and then project.

\[
x_{k+1}^1 = x_k - \eta \nabla f(x_k)
\]

\[
x_{k+1} = \Pi_K(x_{k+1}^1). \quad \text{closet pt.}
\]

Claim: \( \| x_{k+1}^1 - x^* \|^2 \leq \| x_{k+1} - x^* \|^2 \).

If: Pythagoras.

Also: Frank Wolfe algorithm.
Online Convex Opt: ☺

Thm: Fix $\epsilon > 0$. Given a sequence $x_1, x_2, \ldots, x_T$, we play $x_1, x_2, \ldots, x_T$ and receive $f_1(x_1), f_2(x_2), \ldots, f_T(x_T)$. Let $f_t : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex for $\forall x \in \mathbb{R}^d$. Then, for $x^* \in \mathbb{R}^d$,

$$\text{Avg Regret} = \frac{1}{T} \sum_{t=1}^{T} f_t(x_t) \leq \frac{1}{T} \sum_{t=1}^{T} f_t(x^*) + \epsilon$$

after $T \geq \frac{4}{\epsilon^2} \ln \left( \frac{K}{\epsilon} \right)$ steps.

Proof: just put the everywhere instead of $f$.

Hence: get a low regret algorithm with different parameters.

For old case: $k = \Delta_N \Rightarrow D = 1$.

$G = \max_x \| D f(x) \| \leq \sqrt{N}$.

$\Rightarrow$ regret $\leq \epsilon$ after $T \geq \frac{N}{\epsilon^2}$ steps, much weaker. Do better later.

Other Observations: no explicit dependence on $n$ (dimension free).

- Not diff? use subgradients.
- Compute gradient? use stochastic gradient.

For $f(x) = \sum_{t=1}^{T} f_t(x)$, the regret is given by

$$\text{Regret} = \sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \epsilon \sqrt{T} + \epsilon T$$

with high probability.

Varying $\lambda$, high variance. So use carefully. Much work, maybe do later in course.

Some more arguments to speed things up:

1. **Strongly convex**: $f'(y) \geq f'(x) + (f'(x), y-x) + \frac{\lambda}{2} \| y-x \|^2. \Rightarrow H(x) \geq \frac{\lambda}{2}$.

   [Exercise] with a slightly different proof (almost same) get $T = O\left(\frac{1}{\epsilon^2}\right)$ instead, and $k = \frac{1}{\epsilon}$.

   [Also in online] give intuition!!

2. **Smooth**: $f'(x) \leq f'(y) + \frac{\lambda}{2} \| y-x \|^2. \Rightarrow H(x) \leq \frac{\lambda}{2}$. Then get $T = O\left(\frac{1}{\epsilon^2}\right) [\text{but not online convergence}]$.
And suppose:

in "well conditioned", both \( \beta \)-smooth, \( 2 \times \beta \)-strongly convex.

\[
\alpha \frac{\beta}{2} I \leq H F I \leq \beta I
\]

Then \( T = O(\log \frac{1}{\epsilon}) \) suffices! [Linear convergence].

\[\text{pf: } f(x_t + \Delta_t) - f(x_t) \leq \langle \nabla f(x_t), x_t + \Delta_t - x_t \rangle + \frac{\beta}{2} \| x_t + \Delta_t - x_t \|^2 \]

\[
= -\frac{\beta}{2} \| \nabla_t \|^2 + \frac{\beta}{2} \| \nabla_t \|^2 \eta^2
\]

Set \( \eta = \frac{1}{\beta} \) \[\Rightarrow \quad \| \nabla_t \|^2 \leq -\frac{1}{2\beta} \| \nabla_t \|^2. \]

\[f(x_t) - f(x^*) \leq \langle \nabla f(x_t), x_t - x^* \rangle - \frac{\alpha}{2} \| x_t - x^* \|^2 \]

\[\leq \| \nabla_t \| \cdot \| x_t - x^* \| - \frac{\alpha}{2} \| x_t - x^* \|^2. \]

\[\leq \frac{1}{2\alpha} \| \nabla_t \|^2. \]

\[\Rightarrow \quad f(x_{t+1}) - f(x_t) \leq -\frac{\beta}{2} \| \nabla_t \|^2 \leq -\frac{\beta}{2} (f(x_t) - f(x^*)) \quad \text{by \( \beta \)-smooth}. \]

\[\Rightarrow \quad \Delta_t = f(x_t) - f(x^*) \]

\[\Rightarrow \quad \Delta_{t+1} - \Delta_t \leq -\frac{\beta}{2} \Delta_t \quad \Rightarrow \quad \Delta_{t+1} \leq (1 - \frac{\beta}{2}) \Delta_t \]

\[\Rightarrow \quad \Delta_t \leq e^{-\frac{\beta t}{2}} \Delta_0. \]

\[\Rightarrow \text{to be within } \epsilon \text{ of } f(x^*) \text{ need } \frac{B}{\alpha} \log \left( \frac{\Delta_0}{\epsilon} \right) \text{ rounds.} \]

"Linear convergence"

# of rounds increases as \( O(1) \) for each extra bit of accuracy.