Convergent Optimization and First/Second Order Methods.

Convex functions

\[ f: \mathbb{R} \to \mathbb{R} \]

\[ f: \mathbb{R}^2 \to \mathbb{R} \]

Defn. "zeroth order" \( f \) \( x, y \)

\[ f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \]

1st order (if \( f \) is differentiable) \(< \nabla f(x), y - x > \)

\[ \nabla f(x) = \left( \frac{df(x)}{dx_1}, \frac{df(x)}{dx_2}, ..., \frac{df(x)}{dx_n} \right) \]

Assume for all function today
Define the Hessian $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$, where $\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ are second derivatives.

Then $\nabla^2 f(x)$ is positive semi-definite if $x \in$ domain if it has all non-negative eigenvalues.

Example: $f(x) = \|x\|_2^2$.

- Show for scalars: For vectors, do this coordinate by coordinate.

\[ (2a + (1-\lambda)b)^2 \leq 2a^2 + (1-\lambda)b^2 \]

\[ \|a+b\|^2 = \|a\|^2 + \|b\|^2 \]

\[ \|a+b\|^2 = 2a^2 + (1-\lambda)b^2 - 2\lambda a(b - 2a^2 - 2\lambda(b - 2a) = 0 ) \]

\[ \lambda (1-\lambda)a^2 + (1-\lambda)\lambda b^2 - 2\lambda (1-a)b = 0 \]

\[ \lambda (1-\lambda) [a^2 + b^2 - 2ab] \geq 0 \]

\[ 1^{st} \text{ order}: \nabla f(x) = 2x \]

$\exists f(y) = \|y\|^2 \geq \|x\|^2 + \langle 2x, y - x \rangle$?

\[ y^2 - x^2 - 2xy + 2x^2 \geq 0 \]

\[ \geq y^2 + x^2 - 2xy \geq 0 \]

\[ 2^{nd} \text{ order}: \nabla^2 f(x) = 2\text{Identity which is PSD} \]

has all eigenvalues = 1 $\geq 0$
Similarly: \[ f(x) = \sum x_i \log x_i \] Negative entropy.

\[ f(x) = x^TQx + r^Tx \] Quadratic fn.

where \( Q \) is psd

\[ f(x) = \log(1 + e^{-b\langle x, x' \rangle}) \] Logistic loss (max. as log likelihood in logistic regression)

et al.

Linear function! (Reference)

---

Problems to solve:

**Unconstrained Convex Opt**: \[ \min_{x\in\mathbb{R}^n} f(x) \].

**Constrained Convex Opt**: \[ \min_{x\in K} f(x) \] for convex body \( K \subset \mathbb{R}^n \).

Optimality: Convexity: every local optimum is also a global optimum.

Unconstrained: \( x \) is minimizer \( \iff \nabla f(x) = 0 \).

Constrained: \( x \) minimizer \( \iff \langle \nabla f(x), y - x \rangle \geq 0 \) \( \forall y \in K \).

All directions within \( K \) have positive correlation with gradient (i.e., function rising without direction)
So great, solved problem. But no.

1. May not have explicit form for \( f \), so difficult to compute gradient explicitly.

2. Constrained problem, not enough to compute gradients.

3. Even if compute gradient, how to solve \( \nabla f(\mathbf{x}) = 0 \)?

Eq.

\[
f(\mathbf{x}) = \sum_{i=1}^{m} \| b_i - \langle a_i, \mathbf{x} \rangle \|^2
\]

\[
\Rightarrow \nabla f(\mathbf{x}) = \sum_{i} 2 (b_i - \langle a_i, \mathbf{x} \rangle) a_i
\]

\[
= 2A^T(b - Ax)
\]

Set this to zero means.

\[
2A^TAX = A^Tb
\]

Solve a linear system (system of linear equations).

Can use Gaussian elimination.

But just as well can use a convex optimization algorithm, i.e. \( \min \ f(x) \).

See HW for details!!
First Order methods:

Assume can set $\nabla f(x)$ at a point $x$. "Gradient Methods"

Second Order methods

Use both $\nabla^2 f(x)$ and $\nabla f(x)$ Hessian.

typically faster
but require functions to be nicer
and that we start "close" to optimum

Also zeroth Order methods

Only function evaluation is allowed.

(Sometimes can use this to estimate gradient, see SGD.)

---

So Gradient Descent (Cauchy?)

$x_0 \leftarrow$ some starting point
for $t = 0 \ldots T-1$

$x_{t+1} \leftarrow x_t - \eta_t \cdot \nabla f(x_t)$

[Finally: output either $x_T$ or $\hat{x}_T = \left( \sum_{t=1}^{T} x_t \right) / T$]

Facts: if stepsize chosen appropriately, GD converges.

And it converges "fast".

Framework: $\downarrow$ Framework: $\downarrow$ Framework: $\downarrow$ Framework: $\downarrow$
Depends on what assumption we make on the function, etc.

Basic GD Thm: suppose $D$ is the distance $G \to \text{gradient}$

then $\exists \text{ choice of } \eta \text{ s.t. } f(x_T) \leq f(x^*) + \varepsilon.$

$\therefore T = O\left(\frac{G \cdot D}{\varepsilon^2}\right)$ steps for $\varepsilon$ accuracy.

[If imagine $G, D$ "constants", then $O\left(\frac{1}{\varepsilon^2}\right)$ steps for $\varepsilon$ accuracy.]

- We'll prove this theorem, (and its constrained variant)
  - relate it to follow the regularized leader (FTRL)

- Discuss other properties of convex functions that give better convergence

- Stochastic Gradient Descent

- Mirror Descent (aka Preconditioned Gradient Descent)