Lecture 13 Basic Optimization

Friday, October 14, 2016 1:20 PM

Goal: $\min f(x)$ ← objective

$x \in B$ ← constraint

usually $B \subseteq \mathbb{R}^n$, $f: \mathbb{R}^n \to \mathbb{R}$

- Convexity
  - idea: if x, y are valid, $\alpha x + (1-\alpha) y$ is also valid ($\alpha \in [0,1]$)
  - set B is convex, if $\forall x, y \in B$
    $\alpha x + (1-\alpha) y \in B$ ($\alpha \in [0,1]$)
  - function $f(x)$ is convex, if $\forall x, y \in B$
    $f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$
    ($\alpha \in [0,1]$)
  - convex optimization: both $f(x)$ and B are convex

- Basic Algorithm: Gradient Descent
  - recap: Gradient $\nabla f(x) \in \mathbb{R}^n$, 1st order derivative

$$(\nabla f(x))_i = \frac{\partial}{\partial x_i} f(x)$$

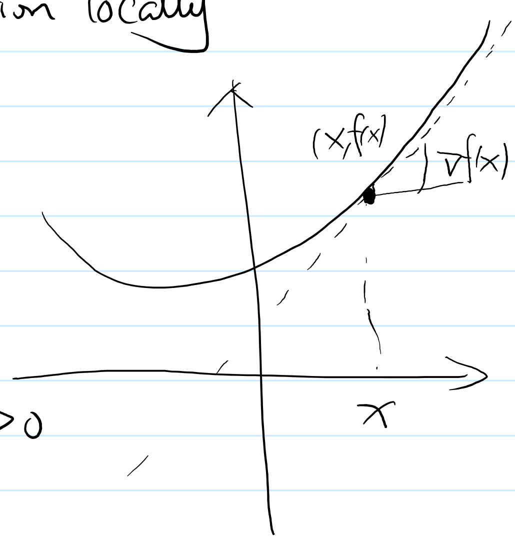
Hessian: 2nd order derivative $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$

$$(\nabla^2 f(x))_{i,j} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$$

- General idea in optimization
  approximate the objective function locally

- if $f(x)$ is convex, then

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$



→ intuition: if $\langle \nabla f(x), y-x \rangle > 0$
$f(y)$ is always worse.
needs an upperbound to guarantee decrease.

- Lipschitz Gradient / "Smoothness"

Def: $f(x)$ is L-smooth (L-Lipschitz Gradient)
if $\forall x, y$ $\quad \|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$

$$\iff f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$$

$$\iff \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \quad (*)$$

- Analyzing Gradient descent

$$\min_y f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$$

solution: $y = x - \frac{1}{L} \nabla f(x)$

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y-x\|^2$$
$$\le f(x) - \frac{1}{2L}\|\nabla f(x)\|^2 \qquad (**)$$

(need to show $\|\nabla f(x)\|$ large to make progress)

$$x^{k+1} = x^k - \eta \nabla f(x^k) \qquad \left(\eta \in \left(0, \tfrac{2}{L}\right]\right)$$

Let $r_k = \|x^k - x^*\|$, first show never gets further

$$r_{k+1}^2 = \|x^k - x^* - \eta \nabla f(x^k)\|^2$$
$$= r_k^2 - 2\eta \langle \nabla f(x^k), x^k - x^* \rangle + \eta^2 \|\nabla f(x^k)\|^2$$

$\downarrow (*), \nabla f(x^*) = 0$

$$\frac{1}{L}\|\nabla f(x^k)\|^2$$

$$\le r_k^2 - \eta\left(\frac{2}{L} - \eta\right)\|\nabla f(x^k)\|^2 \le r_k^2$$

$\Longrightarrow$ always move closer!

Let $\Delta_k = f(x^k) - f(x^*)$, then

$$\Delta_k \underset{\text{convexity}}{\le} \langle \nabla f(x^k), x^k - x^* \rangle \le r_k \|\nabla f(x^k)\|$$
$$\le r_0 \|\nabla f(x^k)\|$$

$$\Delta_{k+1} \underset{\text{same as } (**)}{\le} \Delta_k - \underbrace{\eta\left(1 - \frac{L\eta}{2}\right)}_{\omega} \|\nabla f(x^k)\|^2 \le \Delta_k - \frac{\omega}{r_0^2}\Delta_k^2$$

idea: if $\Delta_k = \frac{r_0^2}{[illegible]}$, $\Delta_{k+1} \le \frac{r_0^2}{[illegible]}$

$w(t+1)$

do induction carefully,

$$f(x^k) - f(x^*) \le \frac{2L\|x^0 - x^*\|^2}{k+4}. \qquad \square$$

Strong convexity: better lowerbound

Def: f is μ-strongly convex if $\forall x, y$

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

Lemma: If f is L-smooth, μ-strongly convex, then

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2$$

Theorem: Choose $\eta = \frac{2}{\mu + L}$, then

$$\|x^k - x^*\| \le \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^k \|x^0 - x^*\|$$

Lecture 14 Stochastic Gradient and Variance Reduction

Sunday, October 16, 2016 10:25 PM

- Stochastic Gradient descent
  - least squares

$$\min \|y - Ax\|^2$$

$$\frac{1}{2n}\sum_{i=1}^{n} (y_i - \langle a_i, x\rangle)^2 \quad , \quad a_i \in R^d$$

For simplicity assume $\|a_i\| = 1$

$$y_i = \langle a_i, x^*\rangle + \varepsilon_i \quad \left(\begin{array}{l}\sum \varepsilon_i a_i = 0 \\ |\varepsilon_i| \le \sigma\end{array}\right)$$

  - can rewrite

$$f(x) = f(x^*) + \frac{1}{2}(x - x^*)^T M (x - x^*)$$

where $M = \frac{1}{n}\sum_{i=1}^{n} a_i a_i^T$

  - SGD for least squares

$$f_i(x) = \frac{1}{2}(y_i - \langle a_i, x\rangle)^2$$

pick random i

$$x^{t+1} = x^t - \eta \nabla f_i(x)$$

$$= x^t + \eta (y_i - \langle a_i, x^t\rangle) a_i$$

  - Analyzing SGD

$$x^{t+1} = x^t - \eta \nabla f_i(x^t) = x^t - \eta(\nabla f(x^t) + \xi_i)$$

ξ_i independent of x^t. $\mathbb{E}[\xi_i] = 0$

Let $r_t = \mathbb{E}[\|x^t - x^*\|^2]$

$$r_{t+1}^2 = r_t^2 - 2\,\mathbb{E}[\eta \langle \nabla f(x^t) + \xi_i, x^t - x^* \rangle] + \eta^2\,\mathbb{E}[\|\nabla f(x^t) + \xi_i\|^2]$$

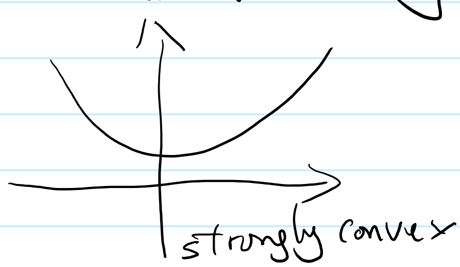
$$= r_t^2 - 2\eta \langle \nabla f(x^t), x^t - x^* \rangle + \eta^2\,\mathbb{E}[(y_i - \langle a_i, x^t \rangle)^2]$$

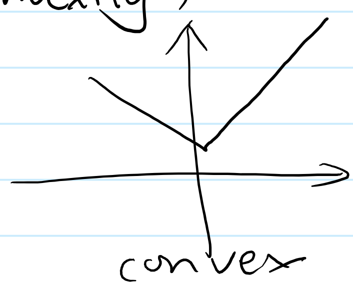
$$= r_t^2 - 2\eta\,(x^t - x^*) M (x^t - x^*) + 2\eta^2 f(x^t)$$

$$= r_t^2 - 2\eta\,(x^t - x^*) M (x^t - x^*) + 2\eta^2 \left(f(x^*) + \frac{1}{2}(x^t - x^*) M (x^t - x^*)\right)$$

- Suppose $\sigma_{\min}(M) = \mu$

(this is called Strong Convexity)





$$r_{t+1}^2 \le r_t^2 - \underbrace{(2\eta - \eta^2)\,\mu\, r_t^2}_{A} + \underbrace{2\eta^2 f(x^*)}_{B}$$

we want term A to dominate term B!

set $\eta \ll \dfrac{\mu r_t^2}{f(x^*)}$ works

in that case

$$r_{t+1}^2 \le r_t^2 (1-\eta)$$
$$\le r_t^2 \left(1 - \frac{\mu r_t^2}{f(x^*)}\right)$$

again we solve the recursion and get

$$r_t^2 = \frac{f(x^*)}{\mu t}$$ (if the initial point is close enough)

in the best case, $\mu = \frac{1}{d}$
(because $tr(M) = \frac{1}{n} \sum \|a_i\|^2 = 1$)
so we can hope to get reasonably close after $O(d)$ iterations.

- System of linear equations

$$r_{t+1}^2 \le r_t^2 - \underbrace{(2\eta - \eta^2)\mu r_t^2}_{A} + \underbrace{2\eta^2 f(x^*)}_{B}$$

we had to use a small η to let A dominate B.
what if $f(x^*) = 0$? (this means $y_i = \langle a_i, x^* \rangle$)
then we can choose $\eta = 1$ and

$$r_{t+1}^2 \le (1-\mu) r_t^2$$
$$\Longrightarrow \quad r_t^2 \le (1-\mu)^t r_0^2$$

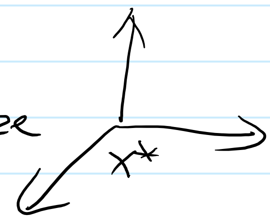
r_t^2 decreases by a constant factor every $\frac{1}{\mu}$ iterations!

- Variance Reduction

- idea: Previously, things did not work because

$$\nabla f_i(x^*) \neq 0$$

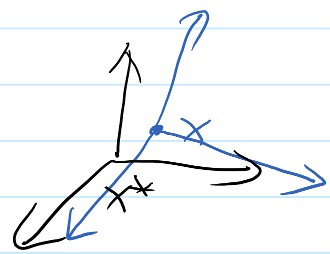
- if we choose a large step size will go away even if we are already at x^* !



- idea: if x is very close to x^*

$$\nabla f_i(x) \approx \nabla f_i(x^*)$$

Fix $\tilde{x}^0 = x$, pick i randomly



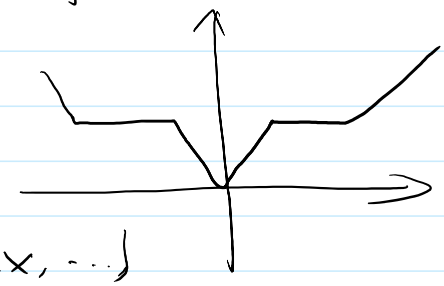
$$\tilde{x}^{t+1} = \tilde{x}^t - \eta\Big(\underbrace{\nabla f_i(\tilde{x}^t) - \nabla f_i(x)}_{\text{variance reduction}} + \nabla f(x)\Big)$$

make sure $\mathbb{E}[\,] = \nabla f(\tilde{x}^t)$

Lecture 15 Non-convex Optimization I Local Analysis

Sunday, October 23, 2016 10:59 PM

- Non-convex optimization
- what can a non-convex function look like?
  - simpler case
    still has a unique minimum.
    (quasi-convex, pseudo-convex, ...)



  - complicated case
    multiple local optima..

- optimality conditions
  - first order optimality condition

    $$\nabla f(x) = 0$$

    - such points are called critical points.
    - for (strongly) convex function, $\nabla f(x) = 0 \Rightarrow x$ is optimal.
  - second order condition

    $$\nabla^2 f(x) \succeq 0$$

    e.g. $f(x) = x_1^2 + x_2^2$

    $$\nabla^2 f(x) = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} \succeq 0.$$

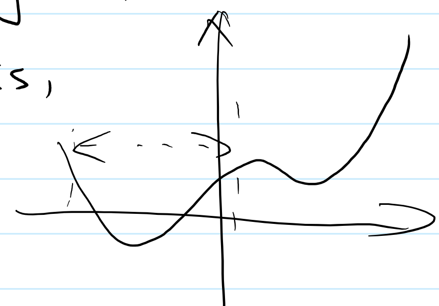
  - saddle points

    $\nabla^2 f(x)$ is not positive semidefinite or negative semidefinite.

e.g. $f(x) = x_1^2 - x_2^2$

$$\nabla^2 f(x) = \begin{pmatrix} 1 & \\ & -1 \end{pmatrix}$$

- Local convergence vs global convergence.
  - when multiple local minima exists, can hope to converge to global minimum with good initialization.



  - local structure of a non-convex function can behave like a convex function!

- Approximate Gradient Descent.
  - idea: after a good initialization, maybe the function is very similar to convex.
  - how to measure "similarity" to convex?
  - Consider gradient descent, initial z^0, goal z^*

    z^* is the optimum for convex function $f(z)$
    however, only has non-convex function $g(z)$
    hope: $g(z)$ close to $f(z)$

$$z^{t+1} = z^t - \eta \nabla g(z^t)$$

- Def: g is $(\alpha, \beta, \varepsilon)$-correlated if

$$\langle \nabla g(z^t), z^t - z^* \rangle \geq \alpha \| z^t - z^* \|^2 + \beta \| \nabla g(z^t) \|^2 - \varepsilon$$

note: if f is μ-strongly convex and L-smooth

$$\langle \nabla f(z^t), z^t - z^* \rangle \geq \frac{\mu L}{\mu + L} \| z^t - z^* \|^2 + \frac{1}{\mu + L} \| \nabla g(z^t) \|^2$$

$$\langle \nabla f(z^t), z^t - z^* \rangle \geq \frac{\mu L}{\mu + L} \|z^t - z^*\|^2 + \frac{1}{\mu + L} \|\nabla g(z^t)\|^2$$

$\left(\frac{\mu L}{\mu + L}, \frac{1}{\mu + L}, 0\right)$ - correlated.

Theorem: $\|z^{t+1} - z^*\|^2 \leq (1 - 2\alpha\eta)\|z^t - z^*\|^2 + 2\eta\varepsilon$, (when $\eta \leq 2\beta$), in particular

$$\|z^t - z^*\|^2 \leq (1 - 2\alpha\eta)^t \|z^0 - z^*\|^2 + \frac{\varepsilon}{\alpha}$$

Proof: $\|z^{t+1} - z^*\|^2 = \|z^t - z^*\|^2 - 2\eta\langle \nabla g(z^t), z^t - z^* \rangle + \eta^2 \|\nabla g(z^t)\|^2$

$$= \|z^t - z^*\|^2 - \eta\left(2\langle \nabla g(z^t), z^t - z^* \rangle - \eta\|\nabla g(z^t)\|^2\right)$$

$$\leq \|z^t - z^*\|^2 - \eta\left(2\alpha\|z^t - z^*\|^2 + (2\beta - \eta)\|\nabla g(z^t)\|^2 - 2\varepsilon\right)$$

$$\leq (1 - 2\alpha\eta)\|z^t - z^*\|^2 + 2\eta\varepsilon$$