Convex Optimization

Prof. Nati Srebro

Lecture 4:

Newton's Method

Reading: Boyd and Vandenberghe 9.5-9.7

Gradient Descent with Backtracking Linesearch

Init	$x^{(0)} \in dom(f)$
Iterate	$\Delta x^{(k)} = -\nabla f\left(x^{(k)}\right)$
	Stop if $\left\|\nabla f\left(x^{(k)}\right)\right\|_2^2 \leq 2\mu\epsilon$
	Set $t^{(k)}$ by backtracking linesearch with params α, β
	$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$

If f is $\boldsymbol{\mu}$-strongly-convex and $\boldsymbol{M}$-smooth on $\left\{f(x) \leq f\left(x^{(0)}\right)\right\}$, then total #func evals and runtime:

$$O\left(\left\lceil \frac{\log M}{\log \frac{1}{\beta}} \right\rceil \frac{1}{\alpha} \max\left(\frac{1}{M}, \frac{1}{\beta}\right) \boldsymbol{\kappa} \log\left(\frac{f\left(x^{(0)}\right) - p^*}{\epsilon}\right)\right)$$

$\boldsymbol{\kappa} = {}^{\boldsymbol{M}}/_{\boldsymbol{\mu}}$

Is **Strong** convexity necessary?

- Can we use Gradient Descent if $f(x)$ is M-smooth, convex, but **not** strongly convex (or with very small μ) ?

- Solution 1: optimize $f_\lambda(x) \stackrel{\text{def}}{=} f(x) + \frac{\lambda}{2}\|x\|^2$
  - $\nabla^2 f_\lambda(x) = \nabla^2 f(x) + \lambda I$
  - f_λ is $(M+\lambda)$-smooth and λ-strongly convex
  - ➔ GD on f_λ finds $f_\lambda\left(x^{(k)}\right) \le \inf_x f_\lambda(x) + \epsilon$ with $k = O\left(\frac{M+\lambda}{\lambda}\log\frac{1}{\epsilon}\right)$

$$f\left(x^{(k)}\right) \le f_\lambda\left(x^{(k)}\right) \le \inf_x f_\lambda(x) + \epsilon \le f_\lambda(x^*) + \epsilon = f(x^*) + \frac{\lambda}{2}\|x^*\|^2 + \epsilon \le p^* + 2\epsilon$$

  - Overall complexity:

$$O\left(\frac{M\|x^*\|_2^2}{\epsilon}\log\frac{1}{\epsilon}\right)$$

$$\boldsymbol{\lambda = \frac{2\epsilon}{\|x^*\|^2}}$$

Is **Strong** convexity necessary?

- Can we use Gradient Descent if $f(x)$ is M-smooth, convex, but **not** strongly convex (or with very small μ) ?

- Solution 1: optimize $f_\lambda(x) \stackrel{\text{def}}{=} f(x) + \frac{\lambda}{2}\|x\|^2$
  - Using $\boldsymbol{\lambda} = \frac{\mathbf{2\epsilon}}{\|\boldsymbol{x}^*\|^2}$, #iter to find ϵ-subopt: $O\left(\frac{M\|x^*\|_2^2}{\boldsymbol{\epsilon}} \log\frac{1}{\epsilon}\right)$
  - Can avoid log-factor by gradually decreasing λ

- Solution 2: Run Gradient Descent directly on $f(x)$, with backtracking linesearch
  - Can find ϵ-suboptimal solution with #iter

$$O\left(\frac{M\left\|x^* - x^{(0)}\right\|_2^2}{\boldsymbol{\epsilon}}\right)$$

Dependence on Conditioning

- Classical analysis depends on smoothness and strong convexity:

$$O\left(\frac{M}{\mu}\log\left(\frac{f\left(x^{(0)}\right)-p^*}{\epsilon}\right)\right)$$

- Linear dependence on M/μ is real, and necessary in order to obtain linear convergence (#iter $\propto \log 1/\epsilon$)
- Can always replace strong convexity $\boldsymbol{\mu}$ with dependence on $\|\boldsymbol{x}^*\|_{\mathbf{2}}^{\mathbf{2}}/\epsilon$

- All quantities depend on choice of norm $\|\cdot\|_{\mathbf{2}}$ and so on choice of basis:

  smoothness $\boldsymbol{M}$, strong convexity $\boldsymbol{\mu}$, $\|\boldsymbol{x}^*\|_{\mathbf{2}}$

$$f(x)+\langle\nabla f(x),\Delta x\rangle+\frac{\boldsymbol{\mu}}{2}\|\boldsymbol{\Delta x}\|_{\mathbf{2}}^{\mathbf{2}} \le f(x+\Delta x) \le f(x)+\langle\nabla f(x),\Delta x\rangle+\frac{\boldsymbol{M}}{2}\|\boldsymbol{\Delta x}\|_{\mathbf{2}}^{\mathbf{2}}$$

$$\boldsymbol{\mu} = \inf_{\|v\|_2=1} v^\top(\nabla^2 f)v \qquad\qquad \boldsymbol{M} = \sup_{\|v\|_2=1} v^\top(\nabla^2 f)v$$

Pre-Conditioned Gradient Descent

Init	$x^{(0)} \in dom(f)$
	$\boldsymbol{H} = \nabla^2 f\left(x^{(0)}\right)$
Iterate	$\Delta x^{(k)} = -\boldsymbol{H}^{-1} \nabla f\left(x^{(k)}\right)$
	Set $t^{(k)}$ by backtracking linesearch with params α, β
	$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$

Complexity depends on conditioning w.r.t. the norm $\|\boldsymbol{x}\|_{\boldsymbol{H}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{x}}$

$$\mu = \inf_{\|\boldsymbol{v}\|_{\boldsymbol{H}}=\mathbf{1}} v^\top \left(\nabla^2 f(x)\right) v \qquad M = \sup_{\|\boldsymbol{v}\|_{\boldsymbol{H}}=\mathbf{1}} v^\top \left(\nabla^2 f(x)\right) v$$

If the Hessian is fixed, $\nabla^2 f(x) = \nabla^2 f\left(x^{(0)}\right) = H$,

$$\mu = \inf_{\|\boldsymbol{v}\|_{\boldsymbol{H}}=\mathbf{1}} v^\top \boldsymbol{H} v = 1 \qquad M = \sup_{\|\boldsymbol{v}\|_{\boldsymbol{H}}=\mathbf{1}} v^\top \boldsymbol{H} v = 1$$

Pre-Conditioning a Quadratic

$$f(x) = \frac{1}{2}x^\top H x + bx \qquad H = \nabla^2 f\left(x^{(0)}\right)$$

- The optimum x^* is given by:

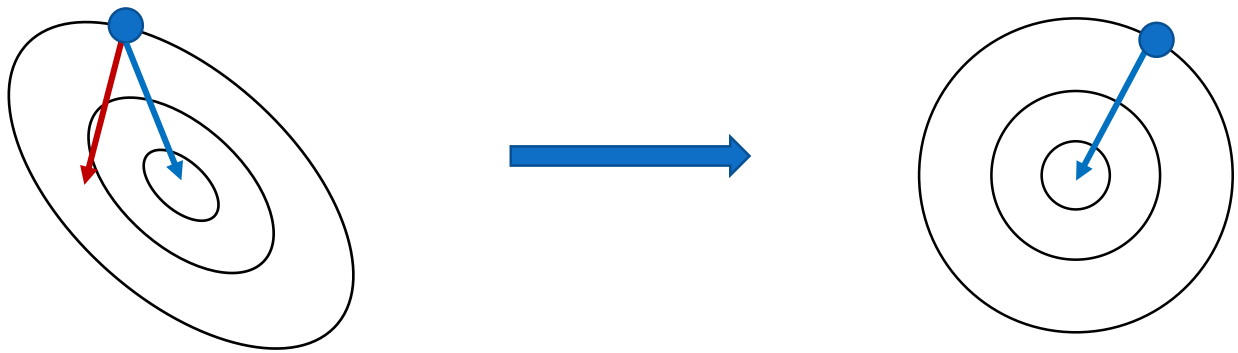
$$0 = \nabla f(x^*) = Hx^* + b \;\rightarrow\; x^* = -H^{-1}b$$

- Taking a single step of preconditioned GD:

$$\Delta x^{(0)} = -H^{-1}\nabla f\left(x^{(0)}\right) = -H^{-1}(Hx + b) = -(x + H^{-1}b)$$

- With a stepsize of $t^{(0)} = 1$:

$$x^{(1)} = x^{(0)} + \Delta x^{(0)} = x^{(0)} - \left(x^{(0)} + H^{-1}b\right) = -H^{-1}b = x^*$$



Newton's Method

$$\min f(x)$$

Access to 1st *and* 2nd order oracles: $x \mapsto f(x), \nabla f(x), \nabla^2 f(x)$

Init	$x^{(0)} \in dom(f)$
Iterate	$\Delta x^{(k)} = -\nabla^2 f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right)$
	Set $t^{(k)}$ by backtracking linesearch with params α, β
	$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$

- Affine invariant: consider $\tilde{f}(\tilde{x}) = f(T\tilde{x} + b)$, and running Newton on $\tilde{f}$ and on f
  - $\Delta x = T \Delta \tilde{x}$
  - If $x^{(0)} = T\tilde{x}^{(0)}$ then $x^{(k)} = T\tilde{x}^{(k)}$

Newton's Method as Minimizing 2nd Order Approximation

$$f\left(x^{(k)} + \Delta x\right) \approx \underbrace{f\left(x^{(k)}\right) + \langle \nabla f\left(x^{(k)}\right), \Delta x\rangle + \frac{1}{2}\Delta x^\top \nabla^2 f\left(x^{(k)}\right)\Delta x}_{\hat{f}^{(k)}\left(x^{(k)} + \Delta x\right)}$$

$$\Delta x^{(k)} = \arg\min_{\Delta x} \hat{f}^{(k)}\left(x^{(k)} + \Delta x\right) = -\nabla^2 f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right)$$

With a step-size of $t^{(k)} = 1$:

$$x^{(k+1)} = \arg\min_{x} \hat{f}^{(k)}(x)$$

Newton's Method

Init	$x^{(0)} \in dom(f)$
Iterate	$\Delta x^{(k)} = -\nabla^2 f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right)$
	Set $t^{(k)}$ by backtracking linesearch with params α, β
	$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$

Stopping condition?

Affine invariant?

Newton's Decrement

$$\hat{f}^{(k)}\left(x^{(k)}+\Delta x\right)=f\left(x^{(k)}\right)+\langle\nabla f\left(x^{(k)}\right),\Delta x\rangle+\frac{1}{2}\Delta x^{\top}\nabla^2 f\left(x^{(k)}\right)\Delta x$$

- Suboptimality according to the quadratic approximation:

$$f^{(k)}\left(x^{(k)}\right)-\min_x \hat{f}^{(k)}(x) \quad = f\left(x^{(k)}\right)-\hat{f}^{(k)}\left(x^{(k)}+\Delta x^{(k)}\right)$$

$$= f\left(x^{(k)}\right)-\left(f\left(x^{(k)}\right)+\langle\nabla f\left(x^{(k)}\right),-\nabla^2 f\left(x^{(k)}\right)\nabla f\left(x^{(k)}\right)\rangle+\frac{1}{2}\left(-\nabla^2 f\left(x^{(k)}\right)^{-1}\nabla f\left(x^{(k)}\right)\right)^{\top}\nabla^2 f\left(x^{(k)}\right)\left(-\nabla^2 f\left(x^{(k)}\right)^{-1}\nabla f\left(x^{(k)}\right)\right)\right)$$

$$=\frac{1}{2}\nabla f\left(x^{(k)}\right)^{\top}\nabla^2 f\left(x^{(k)}\right)^{-1}\nabla f\left(x^{(k)}\right)=\frac{1}{2}\lambda(x)^2$$

- Newton's Decrement: $\lambda(x) \stackrel{\text{def}}{=} \sqrt{\nabla f(x)^{\top}\nabla^2 f(x)^{-1}\nabla f(x)}$

$$\boxed{\text{Stop if } \frac{1}{2}\lambda(x)^2 \leq \epsilon}$$

- **Affine invariant!**

Newton's Method

$$\min f(x)$$

Access to 1st *and* 2nd order oracles: $x \mapsto f(x), \nabla f(x), \nabla^2 f(x)$

Init	$x^{(0)} \in dom(f)$
Iterate	Stop if $\frac{1}{2}\lambda\left(x^{(k)}\right)^2 \leq \epsilon$
	$\Delta x^{(k)} = -\nabla^2 f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right)$
	Set $t^{(k)}$ by backtracking linesearch with params α, β
	$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$

$$\lambda(x) \stackrel{\text{def}}{=} \sqrt{\nabla f(x)^\top \nabla^2 f(x)^{-1} \nabla f(x)}$$

Classical Analysis of Newton's Method

- Assumptions:
  - $\mu I \preccurlyeq \nabla^2 f(x) \preccurlyeq MI$
  - $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x-y\|_2$

- Part I (Damped Phase): $\|\nabla f(x)\| \geq \eta \overset{\text{def}}{=} 3(1-2\alpha)\frac{\mu^2}{L}$

$$f\left(x^{(k)}\right) - f\left(x^{(k+1)}\right) \geq \gamma \overset{\text{def}}{=} \alpha\beta\eta^2 \frac{\mu}{M^2} \quad \Rightarrow k \leq \frac{f\left(x^{(0)}\right) - p^*}{\gamma}$$

  - Backtracking is important! Might take short steps

- Part II (Quadratic Convergence): $\|\nabla f(x)\| < \eta$

$$\left\|\nabla f\left(x^{(k+1)}\right)\right\| \leq \frac{L}{2\mu^2}\left\|\nabla f\left(x^{(k)}\right)\right\|^2$$

  - Stepsize $t = 1$

$$\Rightarrow \left\|\nabla f\left(x^{(k+l)}\right)\right\| \leq \frac{2\mu^2}{L}\Bigg(\overbrace{\frac{L}{2\mu^2}\left\|\nabla f\left(x^{(k)}\right)\right\|}^{\leq \frac{L}{2\mu^2}\eta = 3(1-2\alpha) < \frac{1}{2} \text{ when } \frac{1}{3} \leq \alpha \leq \frac{1}{2}}\Bigg)^{\left(2^l\right)} \leq \frac{2\mu^2}{L} 2^{-2^l}$$

$$\Rightarrow f\left(x^{(k+l)}\right) - p^* \leq \frac{1}{2\mu}\left\|\nabla f\left(x^{(k+l)}\right)\right\|^2 < \frac{2\mu^3}{L^2} 2^{-2^l} \quad \Rightarrow l < \log\log\frac{2\mu^3/L^2}{\epsilon}$$

Classical Analysis of Newton's Method

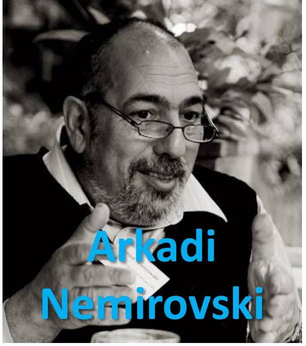
- Assumptions:
  - $\mu I \preccurlyeq \nabla^2 f(x) \preccurlyeq MI$
  - $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\|_2$

Number of iterations to ensure $f\left(x^{(k+l)}\right) \leq p^* + \epsilon$ using $\alpha = 0.4$ and $\beta = 0.9$:

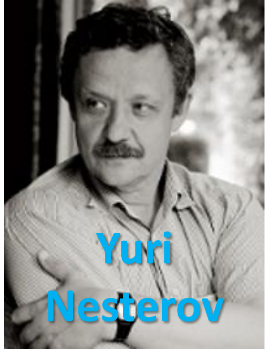
$$k + l \leq \frac{f\left(x^{(0)}\right) - p^*}{0.13\,\frac{\mu^5}{L^2 M^2}} + \log\log\frac{2\mu^3/L^2}{\epsilon}$$

- #grad evals, #Hessian evals: $k + l$
- #func evals: $k \cdot (\#inner\ iter\ of\ backtracking\ linesearch) + l$
- Extra runtime: $O\left((k + l)n^3\right)$

Not affine invariant!



Self Concordant Analysis
(Nemirovski and Nesterov 1994)

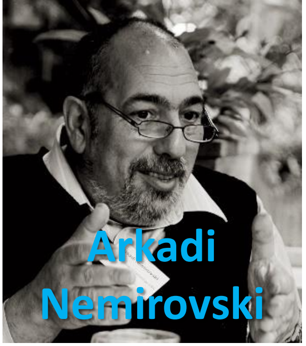


- Def: $f : \mathbb{R}^n \to \mathbb{R}$ is <u>self concordant</u> if
$$\forall_x \forall_{\text{direction } v} |f_v'''(x)| \leq 2 f''(x)^{3/2}$$

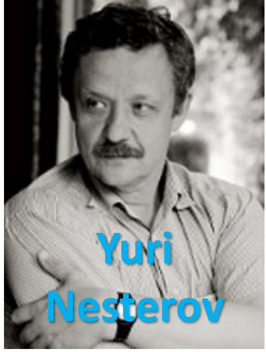
- Examples:
  - Linear ($f_v'''(x) = 0$)
  - Quadratic ($f_v'''(x) = 0$)
  - $f(x) = -\log x$
    - $f'(x) = -\frac{1}{x}, f''(x) = \frac{1}{x^2}, f'''(x) = -\frac{2}{x^3}$
  - What about $f(x) = -\alpha \log x$?
    - We need $\frac{2\alpha}{x^3} \leq 2\frac{\alpha^{3/2}}{x^3}$ ➔ only if $\alpha \geq 1$
  - What about $f(x) = e^x$?
    - $f'(x) = e^x, f''(x) = e^x, f'''(x) = e^x$ ➔ NOT self-concordant

- If $f(x)$ is self-concordant, than so is $\tilde{f}(x) = f(Tx + b)$, for any affine transformation ($T \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$)

  ➔ $f(x) = \log(\langle a, x \rangle + a_0)$ is self concordant



Self Concordant Analysis (Nemirovski and Nesterov 1994)



- Assumption: $|f_v'''(x)| \leq 2f''(x)^{3/2}$

- Suboptimality **bounded** by Newton increment $\lambda(x) = \sqrt{\nabla f(x)^\top \nabla^2 f(x)^{-1} \nabla f(x)}$

$$f(x) \leq p^* + \lambda(x)^2$$

- Part I (Damped Phase): $\lambda(x) \geq \eta \stackrel{\text{def}}{=} \frac{1}{4}(1-2\alpha)$

$$f\left(x^{(k)}\right) - f\left(x^{(k+1)}\right) \geq \gamma \stackrel{\text{def}}{=} \alpha\beta\left(\frac{\eta^2}{1-\eta}\right)$$

  - Backtracking is important! Might take short steps

- Part II (Quadratic Convergence): $\lambda(x) < \eta$

$$\lambda\left(x^{(k+1)}\right) \leq 2\lambda\left(x^{(k)}\right)^2 \qquad \rightarrow f\left(x^{(k+l)}\right) \leq p^* + 2^{-2^l}$$

  - Stepsize $t = 1$

Newton's Method

$$\min f(x)$$

Access to 1st *and* 2nd order oracles: $x \mapsto f(x), \nabla f(x), \nabla^2 f(x)$

Init	$x^{(0)} \in dom(f)$
Iterate	Stop if $\lambda\left(x^{(k)}\right)^2 \leq \epsilon$
	$\Delta x^{(k)} = -\nabla^2 f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right)$
	Set $t^{(k)}$ by backtracking linesearch with params α, β
	$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$

$$\lambda(x) \stackrel{\text{def}}{=} \sqrt{\nabla f(x)^\top \nabla^2 f(x)^{-1} \nabla f(x)}$$

Theorem (NN'94): If $f(x)$ is self-concordant, we have $f\left(x^{(k+l)}\right) \leq p^* + \epsilon$ with

$$k + l \leq \frac{f\left(x^{(0)}\right) - p^*}{\gamma} + \log_2 \log_2 \frac{1}{\epsilon}$$

Depends only on linesearch params
With $\alpha = 0.15, \beta = 0.9: \gamma \approx 1/200$

Newton for Non-Convex Functions

- Is the Newton direction $\Delta x = -\nabla^2 f(x)^{-1} \nabla f(x)$ a descent direction?

$$\langle \nabla f(x), \Delta x \rangle = \langle \nabla f(x), -\nabla^2 f(x)^{-1} \nabla f(x) \rangle = -v^\top \nabla^2 f(x)^{-1} v < 0$$

$v = \nabla f(x)$

If f convex
i.e. $\nabla^2 f \succcurlyeq 0$

- If $f(x)$ is non-convex, Δx might not be a descent direction!

- Newton's method solves $\nabla \hat{f}(x + \Delta x) = 0$.
- $f(x)$ convex ➔ $\nabla^2 f(x) \succcurlyeq 0$ ➔ $\hat{f}(x)$ convex ➔ $\nabla \hat{f} = 0$ is a minimum
- But if $f(x)$ is not convex, $x + \Delta x$ could be a saddle point or even maximum of $\hat{f}$!
- Newton (at least with $t = 1$) converges to nearby critical point of $f(x)$ – local minima, local maxima or saddle point

Newton vs Gradient Descent

Gradient Descent

- Depends on choice of basis
- #iter $\propto \kappa$
- Linear convergence,
$$\propto \log \frac{1}{\epsilon}$$
- 1st order oracle
- Just vector ops—
$O(n)$ time per iter

Newton

- Affine invariant
- Bad conditioning OK
- Quadratic convergence,
$$\propto \log \log \frac{1}{\epsilon}$$
- 2nd order oracle
- Matrix inversion—
$O(n^3)$ per iter