# Smoother Unconstrained Multivariate Optimization

Antonio Frangioni

Department of Computer Science
University of Pisa
https://www.di.unipi.it/~frangio
mailto:frangio@di.unipi.it

Optimization Methods and Game Theory
Master in Artificial Intelligence and Data Engineering
University of Pisa

A.Y. 2022/23

## Outline

## Outline

- Outstanding assumption so far: $d^i = -\nabla f(x^i)$: really needed?
- Crucial convergence arguments:
  1. $\varphi_i'(0) = -\|\nabla f(x^i)\|^2$: "far from $x_*$ the derivative is very negative"
  2. "you can get a non-vanishing fraction of the descent promised by $\varphi_i'(0)$"

     $\equiv$ "exact" LS or Armijo or FS + $L$-smooth $\Longrightarrow$ $\alpha_i$ does not $\rightarrow 0$ "too fast"

  $\Longrightarrow$ "significant decrease at each step unless $\|\nabla f(x^i)\| \rightarrow 0$"
- 2. does not really depend on the chosen direction, and

  $\exists$ many other directions that ensure 1. holds (within some factor)

- Outstanding assumption so far: $d^i = -\nabla f(x^i)$: really needed?
- Crucial convergence arguments:
  1. $\varphi_i'(0) = -\|\nabla f(x^i)\|^2$: "far from $x_*$ the derivative is very negative"
  2. "you can get a non-vanishing fraction of the descent promised by $\varphi_i'(0)$"

     $\equiv$ "exact" LS or Armijo or FS + $L$-smooth $\Longrightarrow$ $\alpha_i$ does not $\to 0$ "too fast"

  $\Longrightarrow$ "significant decrease at each step unless $\|\nabla f(x^i)\| \to 0$"
- 2. does not really depend on the chosen direction, and

  $\exists$ many other directions that ensure 1. holds (within some factor)
- The (parodied) twisted gradient algorithm: "$d^i = -\nabla f(x^i)$ rotated by $\pi / 4$"

  $\equiv$ $d^i = R(-\nabla f(x^i))$, rotation matrix $R$ [7]
- Gives $\varphi_i'(0) = -\|\nabla f(x^i)\|^2 \cos(\pi / 4) < 0$ (**check**)

  $\Longrightarrow$ convergence proofs carry forward largely unchanged
- Not just $\pi / 4$: $\theta$ not too close to $\pi / 2$ $\equiv$ $\cos(\theta)$ "not too small"
- $\infty$-ly many feasible $\theta$ and $\infty$-ly many $\neq d$ for each $\theta$ $\equiv$ $\infty$-ly many $R$

- Descent direction $\equiv \frac{\partial f}{\partial d^i}(x^i) < 0 \equiv \langle d^i , \nabla f(x^i) \rangle < 0 \equiv \cos(\theta^i) < 0$
  $\equiv$ "$d^i$ points roughly in the same direction as $-\nabla f(x^i)$"
- There is a whole half space of descent directions $\implies$ a lot of flexibility
- Zoutendijk's Theorem [3, Th. 3.2]: $f \in C^1$, $f$ $L$-smooth, $f_* > -\infty$,
  $(A) \cap (W) \implies \sum_{i=1}^{\infty} \cos^2(\theta^i) \| \nabla f(x^i) \|^2 < \infty$
- Consequence: $\sum_{i=1}^{\infty} \cos^2(\theta^i) = \infty \implies \{ \| \nabla f(x^i) \| \} \to 0$
  $\equiv$ $d^i$ does not get $\perp \nabla f(x^i)$ "too fast" $\implies$ convergence
- Simple case: $\cos(\theta^i) \geq \bar{\theta} > 0$ (bounded away from 0),
  gradient method just the obvious case, $cos^2(\theta^i) = 1$
- Very many $d^i$ to choose from, but which $d^i$ is better than $-\nabla f$?
- Not clear if you only look to first-order model $\implies$ have to look farther

## Outline

- Want a better direction = faster convergence? Use a better model!
- Next better model to linear ($\equiv$ gradient): quadratic
- $\nabla^2 f(x^i) \succ 0 \implies \exists$ minimum of second-order model $Q_{x^i}(z) \implies$ Newton's direction $d^i = -[\nabla^2 f(x^i)]^{-1} \nabla f(x^i)$ (**check**)
- No problem with the step here, $\alpha^i = 1$ (the minimum $\exists$) $\implies$ Newton's method: $x^{i+1} = x^i + d^i$ (just do step $\alpha^i = 1$ along $d^i$)
- Nonlinear equation interpretation: want to solve $\nabla f(x) = 0$, write $\nabla f(x) \approx \nabla f(x^i) + \nabla^2 f(x^i)(x - x^i)$ and solve linear equation instead
- Newton's not globally convergent $\implies$ has to be globalised
- "Easy" as $\nabla^2 f(x^i) \succ 0 \implies [\nabla^2 f(x^i)]^{-1} \succ 0 \implies d^i$ is of descent:

  $\langle \nabla f(x^i), d^i \rangle = -\nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) < 0$

  (but $< 0$ is not enough, we need it to be "negative enough")

- Globalised Newton's: simply add AWLS / BLS with $\alpha^0 = 1$
- Theorem 1: $f \in C^2$ $L$-smooth and $\tau$-convex $\Longrightarrow$ $\cos(\theta^i) \leq -\tau / L \, [< 0]$ $\Longrightarrow$ global convergence (via Zoutendijk)
- Theorem 2: $f \in C^3$ , $\nabla f(x_*) = 0$ , $\nabla^2 f(x_*) \succ 0 \Longrightarrow \exists\, \delta > 0$ s.t. $x^0 \in \mathcal{B}(x_*, \delta) \Longrightarrow$ "pure" Newton's ($\alpha^i = 1$) $\{x^i\} \to x_*$ quadratically
- Theorem 3: If $\{x^i\} \to x_*$, $\exists\, h$ s.t. $\alpha^i = 1$ satisfies (A) for all $i \geq h$ (requires $m_1 \leq 1/2$, $m_1 > 1/2$ cuts away minimum when $f$ quadratic)
- "Global phase" ($\alpha^i$ varies) + quadratically convergent "pure Newton's phase"
- Pure Newton's phase ends in $O(1)$ ($\approx 6$) iterations in practice
- If $\nabla^2 f$ $M$-smooth then global phase also "$O(1)$" [2, (9.40)]:
  $O(M^2 L^2 (f(x^0) - f_*) / \tau^5)$ (??, but quite fast in practice)

- Theorem 1, two technical steps using $\nabla^2 f(x^i)d^i = -\nabla f(x^i)$:
  - $\langle \nabla f(x^i), d^i \rangle = -(d^i)^T \nabla^2 f(x^i) d^i \leq -\tau \| d^i \|^2$
  - $\| \nabla f(x^i) \| = \| \nabla^2 f(x^i) d^i \| \leq \| \nabla^2 f(x^i) \| \| d^i \| \leq L \| d^i \|$

  $\Longrightarrow \cos(\theta^i) = \langle \nabla f(x^i), d^i \rangle / ( \| \nabla f(x^i) \| \| d^i \| ) \leq -\tau / L$
- Theorem 2: [3, Th. 3.5]
- Theorem 3 (sketch): $\{ x^i \} \to x_* \implies \| \nabla f(x^i) \| \to 0 \implies \| d^i \| \to 0$

$$
\begin{aligned}
f(x^i + d^i) &= f(x^i) + \langle \nabla f(x^i), d^i \rangle + \tfrac{1}{2}(d^i)^T [\nabla^2 f(x^i)] d^i + R(d^i) \\
&= f(x^i) - \nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) \\
&\qquad + \tfrac{1}{2} \nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) + R(d^i) \\
&= f(x^i) - \tfrac{1}{2} \nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) + R(d^i) \\
&= f(x^i) + \tfrac{1}{2} \langle \nabla f(x^i), d^i \rangle + R(d^i)
\end{aligned}
$$

$\varphi'_{x^i,d^i}(0) = \langle \nabla f(x^i), d^i \rangle \to 0$ as $d^i \to 0$, but $R(d^i) \to 0$ faster

$\Longrightarrow$ eventually $R()$ negligible $\Longrightarrow$ eventually (A) holds with $m_1 < 1/2$

**Exercise:** complete the sketch of the proof of Theorem 3

- Interesting interpretation: Newton $\equiv$ Gradient in a twisted space
- $f(x) = \frac{1}{2}x^T Q x + qx$, $d = -x - Q^{-1}q \implies$
  $\nabla f(x + d) = 0$: Newton ends in one iteration
- Relevant object: $Q \succeq 0 \implies Q = RR$, $R = Q^{1/2}$ square root of $Q$
  $\exists$, not unique: a symmetric one is $Q = H \Lambda H^T \implies R = H\sqrt{\Lambda}H^T$ (**check**)
- $Q$ nonsingular $\implies$ $R$ nonsingular $\implies$ $z = Rx \equiv x = R^{-1}z$, a bijection
- $h(z) = f(R^{-1}z) = \frac{1}{2}z^T I z + qR^{-1}z$:
  "in $z$-space, $\nabla^2 f(z^i)$ looks like $I$" $\implies$ gradient is fast
- In fact: $g = -\nabla h(z) = -z - R^{-1}q \implies \nabla h(z + g) = 0$ (**check**)
- Translate $g$ from $z$-space to $x$-space:
  $R^{-1}z = R^{-1}(-z - R^{-1}q) = -z - Q^{-1}q = d$
- $z = Rx$ not the only choice, $z \approx Rx$ ("very $\approx$") works (will see)

- Newton's method $\equiv$ space dilation: a linear map making $\nabla^2 f$ "simple"
- Must it necessarily be $\nabla^2 f(\, x^i \,)^{-1}$? No, especially if $\nabla^2 f(\, x^i \,)^{-1} \not\succeq 0$
- $d^i \leftarrow -[\, H^i \,]^{-1} \nabla f(\, x^i \,)$ , $\tau I \preceq H^i \preceq LI$ , (A) $\cap$ (W) $\Longrightarrow$ global convergence (rewrite Theorem 1 with $H^i$ in place of $\nabla^2 f(\, x^i \,)$)
- $\nabla^2 f \not\succ 0$: choose "small" $\varepsilon^i$ s.t. $H^i = \nabla^2 f(\, x^i \,) + \varepsilon^i I \succ 0$
- Any $\varepsilon^i > -\lambda^n$ works ($\lambda^n < 0$), but numerical issues:
  any double $\leq$ 1e−16 "is 0" (1e−16 very optimistic, at least 1e−12)
- Algorithmic issues: $\lambda^n(\, \nabla^2 f(\, x^i \,) + \varepsilon I)$ "very small" $\Longrightarrow$ axes of $S(\, Q_{x^i} \,, \cdot \,)$ "very elongated" $\Longrightarrow$ "$x^{i+1}$ very far from $x^i$", not good for a local model
- Simple form: $\varepsilon = \max\{\, 0 \,, \, \delta - \lambda^n \,\}$ for appropriately chosen smallish $\delta$ (1e−8? 1e−4? 1e−12? hard to say in general)

- Turns out $\varepsilon = \max\{\, 0 \,,\, \delta - \lambda^n \,\}$ solves $\min\{\, \| \, H - \nabla^2 f(\, x^i \,) \, \|_2 \; : \; H \succeq \delta I \,\}$
- Can use $\neq$ norms: to solve $\min\{\, \| \, H - \nabla^2 f(\, x^i \,) \, \|_F \; : \; H \succeq \delta I \,\}$
  - compute spectral decomposition $\nabla^2 f(\, x^i \,) = H \Lambda H^T$
  - $H^i = H \bar{\Lambda} H^T$ with $\bar{\gamma}^i = \max\{\, \lambda^i \,,\, \delta \,\}$
- In both cases, if $\{\, x^i \,\} \to x_*$ with $\nabla^2 f(\, x_* \,) \succeq \delta I \implies \varepsilon^i = 0 \equiv$ $H^i = \nabla^2 f(\, x^i \,)$ eventually $\implies$ quadratic convergence in the tail
- In both cases, $O(\, n^3 \,)$; say, compute $\lambda^n$ + Cholesky factorization $H^i = L^i (L^i)^T$, $L^i$ triangular (fastest and more stable way)
- Can modify factorization on the fly (diagonal $< 0 \implies$ increase $\varepsilon$) [3, p. 52+]
- Whatever you do, $O(\, n^3 \,)$ too much for large-scale ($n = 10^{4+}$): something way cheaper needed, $O(\, n^2 \,)$ or less

- $\nabla^2 f(x^i) \not\succeq 0 \implies \exists$ negative curvature direction along which $f$ decreases

- $\nabla^2 f(x^i) \not\succeq 0 \implies \exists$ negative curvature direction along which $f$ decreases
  $\equiv$ exactly what we want when minimizing $f$, why excluding them?

- How? $Q_{x^i}(z)$ has no minimum ... on $\mathbb{R}^n$, but it does on a compact set

- $\mathbb{R}^n \supset \mathcal{T}^i =$ (compact) trust region around $x^i$ "where $Q_{x^i}$ can be trusted"
  $x^{i+1} \in \operatorname{argmin}\{ Q_{x^i}(z) : z \in \mathcal{T}^i \}$ a constrained problem

- Even worse: it is $\mathcal{NP}$-hard even for simple $\mathcal{T}$ like $\mathcal{B}_1(x^i, r)$ or $\mathcal{B}_\infty(x^i, r)$

- $\nabla^2 f(x^i) \not\succeq 0 \implies \exists$ negative curvature direction along which $f$ decreases
  $\equiv$ exactly what we want when minimizing $f$, why excluding them?

- How? $Q_{x^i}(z)$ has no minimum …on $\mathbb{R}^n$, but it does on a compact set

- $\mathbb{R}^n \supset \mathcal{T}^i =$ (compact) trust region around $x^i$ "where $Q_{x^i}$ can be trusted"
  $x^{i+1} \in \operatorname{argmin}\{\, Q_{x^i}(z) \,:\, z \in \mathcal{T}^i \,\}$ a constrained problem

- Even worse: it is $\mathcal{NP}$-hard even for simple $\mathcal{T}$ like $\mathcal{B}_1(x^i, r)$ or $\mathcal{B}_\infty(x^i, r)$
  …but not for $\mathcal{B}_2(x^i, r)$: "round balls are simpler than kinky balls"

- An optimization problem with quadratic constraints

- Which $r$?

- Can use any $H^i \approx \nabla^2 f(x^i)$, not necessarily $\succ 0$
- $x^{i+1}$ optimal $\equiv$ $x^{i+1} = x^i + d^i$ and $\exists\, \lambda^i \geq 0$ s.t.

| | | |
|---|---|---|
| $[\, H^i + \lambda^i I \,] d^i = -\nabla f(x^i)$ | [linear] | Karush- |
| $H^i + \lambda^i I \succeq 0$ | [semidefinite] | Khun- |
| $\lambda^i (\, r - \| d^i \| \,) = 0$ | [nonlinear] | Tucker |

- $\lambda > 0 \implies$ like in line search with $\varepsilon^i = \lambda$ (but here $\lambda$ unknown)
- $\| d^i \| < r \implies \lambda^i = 0 \implies$ normal Newton step ($\mathcal{T}$ has no effect)
- $\{ x^i \} \to x_* \implies \| d^i \| \to 0 \implies$ eventually $\lambda^i = 0 \implies$ quadratic convergence in the tail
- Plenty of smart ways to find $\lambda$, $x^{i+1}$ or approximate them (just as well), [3, §4.1], but matrix factorizations may be needed $\implies O(n^3)$ again
- LS: first $d^i$, then $\alpha^i$; TR: first $r\,(\approx \alpha^i)$, then $d^i$. Ultimately, similar
- In both cases, properly choose $H^i \approx \nabla^2 f(x^i)$ to reduce the cost crucial

## Outline

- The space is big

- The space of $H^i$ that give "fast convergence" is big
- Superlinear convergence if "$H^i$ looks like $\nabla^2 f(x^i)$ along $d^i$" [3, Th. 3.6]
  $\lim_{i \to \infty} \| (H^i - \nabla^2 f(x^i)) d^i \| / \| d^i \| = 0$ (don't care elsewhere)
- General derivation of Quasi-Newton methods:
  $$m^i(x) = \nabla f(x^i)(x - x^i) + \tfrac{1}{2}(x - x^i)^T H^i (x - x^i), \; x^{i+1} = x^i + \alpha^i d^i$$
- Having computed $x^{i+1}$ and $\nabla f(x^{i+1})$, new model
  $$m^{i+1}(x) = \nabla f(x^{i+1})(x - x^{i+1}) + \tfrac{1}{2}(x - x^{i+1})^T H^{i+1}(x - x^{i+1})$$
- Nice properties we would like $H^{i+1}$ to have:
  i) $H^{i+1} \succ 0$ (the new model is strongly convex)
  ii) $\nabla m^{i+1}(x^i) = \nabla f(x^i)$ (the new model agrees with old information)
  iii) $\| H^{i+1} - H^i \|$ "small" (the new model is not too different)
- ii) $\equiv$ $H^{i+1}(x^{i+1} - x^i) = \nabla f(x^{i+1}) - \nabla f(x^i)$ "secant equation"

- Depending on choices at iteration $i$, i) $\cap$ ii) may not be possible
- Notation: $s^i = x^{i+1} - x^i = \alpha^i d^i$ , $y^i = \nabla f(x^{i+1}) - \nabla f(x^i)$ (fixed)
  secant equation $\equiv$ (S) $H^{i+1} s^i = y^i$ (**check**)
- (S) $\Longrightarrow \langle s^i , y^i \rangle = (s^i)^T H^{i+1} s^i$ , i) $\cap$ ii) $\Longrightarrow \langle s^i , y^i \rangle > 0$
  "curvature condition" (C), (most often written $\rho^i = 1 \,/\, \langle s^i , y^i \rangle > 0$
- $s^i$ need be properly chosen at iteration $i$ for things to work at $i+1$
- Quasi-Newton $\Longrightarrow$ $d^i$ fixed, but $s^i$ also depends on $\alpha^i$ which is "free"
- Very good news: (W) $\Longrightarrow$ (C)
  Proof: $\varphi'(\alpha^i) = \langle \nabla f(x^{i+1}) , d^i \rangle \geq m_3 \varphi'(0) = m_3 \langle \nabla f(x^i) , d^i \rangle \Longrightarrow$
  $\langle \nabla f(x^{i+1}) - \nabla f(x^i) , d^i \rangle = \langle y^i , d^i \rangle \geq (m_3 - 1) \varphi'(0) > 0$
  but $s^i = \alpha^i d^i$ and $\alpha^i > 0 \Longrightarrow \langle y^i , s^i \rangle = \alpha^i \langle y^i , d^i \rangle > 0$
- Assuming an AWLS, (C) can always be satisfied

- i) $\cup$ i)) $\cup$ iii) $\equiv$ $H^{i+1} = \text{argmin}\,\{\, \| H - H^i \| \;:\; (\text{S})\,,\; H \succeq 0 \,\}$
- Appropriate "$\|\cdot\|$" [3, p. 138]: Davidon-Fletcher-Powell formula

  (DFP) $\quad H^{i+1} = (\, I - \rho^i y^i (s^i)^T \,) H^i (\, I - \rho^i s^i (y^i)^T \,) + \rho^i y^i (y^i)^T$
- $H^{i+1}$ = rank-two correction of $H^i$, $O(\, n^2 \,)$ to produce $H^{i+1}$ out of $H^i$
- Actually need $B^{i+1} = [\, H^{i+1} \,]^{-1}$: Sherman-Morrison-Woodbury formula [8]

  (SMW) $\quad [\, A + ab^T \,]^{-1} = A^{-1} - A^{-1}ab^T A^{-1} \,/\, (\, 1 - b^T A^{-1} a \,)$

  $\Longrightarrow$ (DFP$^{-1}$) $\quad B^{i+1} = B^i + \rho^i s^i (s^i)^T - B^i y^i (y^i)^T B^i \,/\, (y^i)^T B^i y^i$

  $\Longrightarrow$ $O(\, n^2 \,)$ per iteration, just matrix-vector products, no inverse
- This $\approx$ learning $\nabla^2 f$ out of samples of $\nabla f$ (learning2optimize)
- Quite efficient, but can do better

- Write (S) for $B^{i+1}$: $s^i = B^{i+1} y^i \implies B^{i+1} = \operatorname{argmin} \{ \| B - B^i \| : \ldots \}$
  everything is symmetric, just $B \longleftrightarrow H$ and $s \longleftrightarrow y$
- Broyden-Fletcher-Goldfarb-Shanno formulæ [3, p. 139], still $O(\, n^2\,)$:

$$\text{(BFGS)} \quad H^{i+1} = H^i + \rho^i y^i (y^i)^T - H^i s^i (s^i)^T H^i \,/\, (s^i)^T H^i s^i$$

$$\text{(BFGS)} \quad B^{i+1} = (\, I - \rho^i s^i (y^i)^T \,) B^i (\, I - \rho^i y^i (s^i)^T \,) + \rho^i s^i (s^i)^T$$

$$= B^i + \rho^i [\, (\, 1 + \rho^i (y^i)^T B^i y^i \,) s^i (s^i)^T - (\, B^i y^i (s^i)^T + s^i (y^i)^T B^i \,) \,]$$

- Broyden family [3, § 6.3]: $H^{i+1} = \beta H^{i+1}_{\text{DFP}} + (1 - \beta) H^{i+1}_{\text{BFGS}}$, still $O(\, n^2\,)$.
- Surely satisfies (S) and $H^{i+1} \succeq 0$ if $\beta \in [\, 0 \,, 1\,]$ (but $\exists$ feasible $\beta \notin [\, 0 \,, 1\,]$)
- Flexible, good compromise between iteration cost and convergence speed, convergence theory available [3, § 6.4] (not exactly trivial)
- Important choice: $B^0$. Obvious solution $B^0 = \delta I$, but which $\delta$?
  Alternative: $B^0 =$ finite difference $\approx [\, \nabla^2 f(\, x^0\,)\,]^{-1}$

**Exercise:** Discuss how to compute a "finite difference" and how much does it cost

- For very large $n$ even $O(\,n^2\,)$ is way too much
- $O(\,n\,)$ new information per iteration $\nabla f(\,x^i\,)$: only keep/use last $k \ll n$
- Limited-memory BFGS (L-BFGS): just unfold the last $k$ iterations

$$B^{i+1} = (V^i)^T B^i V + \rho^i s^i (s^i)^T \quad \text{with } V^k = I - \rho^i y^i (s^i)^T \quad \equiv$$

$$\begin{aligned} B^{i+1} = {} & (V^{i-k} V^{i-k+1} \ldots V^i)^T B^{i-k} (V^{i-k} V^{i-k+1} \ldots V^i) + \\ & + \rho^{i-k} (V^{i-k+1} \ldots V^i)^T s^{i-k} (s^{i-k})^T (V^{i-k+1} \ldots V^i) + \\ & + \rho^{i-k+1} (V^{i-k+2} \ldots V^i)^T s^{i-k+1} (s^{i-k+1})^T (V^{i-k+2} \ldots V^i) + \\ & + \ldots + \rho^i s^i (s^i)^T \end{aligned}$$

- Memory/time cost per iteration is $O(\,kn\,)$ [3, Algorithm 7.4], but trade-off: convergence worsens as $k \searrow$ ($k$ large $\approx$ Newton but $k$ small $\approx$ gradient)
- Funny tidbit: can choose $B^{i-k}$ arbitrarily anew at each $i$, but of course it need be sparse, e.g., $B^{i-k} = \gamma^i I$ with $\gamma^i = \langle\, s^i\,,\, y^{i-1}\,\rangle \,/\, \|\, y^{i-1}\,\|^2$
- Just one of many possible large-scale quasi-Newton variants [3, Chapter 7]

## Outline

- Twisting $\equiv\ d^i = H^i(\,-\nabla f(\,x^i\,)\,)$ is at least $O(\,n^2\,)$ by definition (not even counting forming $H^i$) unless $H^i$ "very special" $\equiv$ rather dirty tricks
- Cheaper alternative: deflecting $\equiv\ d^i = -\nabla f(\,x^i\,) + v^i$, $O(\,n\,)$ by definition
- But how to choose $v^i$ in the whole of $\mathbb{R}^n$ (cheaply)?
- Simple idea: $v^i = \beta^i d^{i-1}$, direction at previous iteration scaled by some $\beta^i$ (?)
- If $v^0 = 0$, then $d^i = -[\,\sum_{h=1}^{i} \gamma^h \nabla f(\,x^h\,)\,]$ for some $\gamma^i$: (opposite of) aggregated of all past gradients $\equiv$ "history" of computation ($\approx H^i$ in BFGS)
- For twisting, easy to ensure $\varphi'_{x^i,d^i}(\,0\,) < 0$ (just $H^i \succeq 0$) nontrivial to choose $\beta^i$ that does the same (crucial …or not?)
- Will clearly happen as $\beta^i \to 0$ (**check**), but then $d^i \to -\nabla f(\,x^i\,) \implies$ slow
- Need better ideas

## Outline

- Gradient method + exact LS $\Longrightarrow \langle \nabla f(x^{i+1}), d^i \rangle = 0 \equiv d^{i+1} \perp d^i$
  $\equiv x^{i+1}$ minimum over all the (small) subspace spanned by $d^i$
- Property is lost at $i+2$: $x^{i+2}$ not the minimum over $d^i \Longrightarrow$ zig-zags
- Would be nice if $x^{i+1}$ minimum on the subspace spanned by $\{ d^1 , d^2 , \dots , d^i \}$ (getting larger with every iteration)
- Possible for quadratic $f \equiv$ linear systems with $d^i = -\nabla f(x^i) + \beta^i d^{i-1}$
- Requires two conditions (proofs @Federico):
  1. $\beta^i = (\nabla f(x^i)^T Q d^{i-1}) / ((d^{i-1})^T Q d^{i-1})$ a.k.a. Fletcher-Reeves formula
  2. the optimal step is always taken along each $d^i$

  $\Longrightarrow$ all directions are $Q$-conjugate $\equiv (d^i)^T Q d^j = 0 \quad \forall i, j$
  $\Longrightarrow$ the algorithm terminates in at most $n$ iterations (in exact arithmetic)
- Important: F-R formula can be rewritten $\beta^i_{FR} = \| \nabla f(x^i) \|^2 / \| \nabla f(x^{i-1}) \|^2$
  i.e., without any reference to $Q$ , $q \Longrightarrow$ works for any $f(\cdot)$

**procedure** $x = NCG(f, x, \varepsilon)$
  $\nabla f^- = 0$;
  **while**( $\| \nabla f(x) \| > \varepsilon$ ) **do**
    **if**( $\nabla f^- = 0$ ) **then** $d \leftarrow -\nabla f(x)$;
      **else** $\{ \beta = \| \nabla f(x) \|^2 / \| \nabla f^- \|^2; d \leftarrow -\nabla f(x) + \beta d^-; \}$
    $\alpha \leftarrow \text{LS}(f, x, d)$; $x \leftarrow x + \alpha d$; $d^- \leftarrow d$; $\nabla f^- \leftarrow \nabla f(x)$;

- $f$ quadratic + exact LS $\Longrightarrow$ quadratic CG:
  $\nabla f(x) = 0 \equiv Qx = -q$ in at most $n$ iterations (exact arithmetic)
  $\ll n$ iterations if clustered eigenvalues ... (e.g., properly preconditioned)
- Many $\neq$ $\beta$-formulæ, all $\equiv$ for quadratic $f$ but not so here
  1. Polak-Ribière: $\beta^i_{PR} = \langle \nabla f(x^i) - \nabla f(x^{i-1}), \nabla f(x^i) \rangle / \| \nabla f(x^{i-1}) \|^2$
  2. Hestenes-Stiefel:
     $\beta^i_{HS} = \langle \nabla f(x^i) - \nabla f(x^{i-1}), \nabla f(x^i) \rangle / \langle \nabla f(x^i) - \nabla f(x^{i-1}), d^{i-1} \rangle$
  3. Dai-Yuan: $\beta^i_{DY} = \| \nabla f(x^i) \|^2 / \langle \nabla f(x^i) - \nabla f(x^{i-1}), d^{i-1} \rangle$
  4. ...
- LS only exact with quadratic $f$, otherwise AWLS

- Convergence nontrivial, depends a lot on $\beta$-formula + conditions
- F-R requires $m_1 < m_2 < 1/2$ for (A) $\cap$ (W') to work
- (A) $\cap$ (W') $\not\Rightarrow$ $d^i$ of P-R is of descent, unless $\beta^i_{PR+} = \max\{\, \beta^i_{PR} \,,\, 0 \,\}$
  similar $\beta^i_{HS+} = \max\{\, \beta^i_{HS} \,,\, 0 \,\}$ useful for H-S
- The above is a restart: from time to time, take "plain" $-\nabla f$
- Turns out restarts are a good idea, especially for F-R:
  $\|\, \nabla f(\, x^i \,) \,\| \ll \|\, d^i \,\| \iff \cos(\, \theta^i \,) \approx 0 \;\equiv\; \nabla f(\, x^i \,) \approx\perp d^i$
  $\Longrightarrow\; x^{i+1} \approx x^i \;\Longrightarrow\; \cos(\, \theta^{i+1} \,) \approx 0$
  $\Longrightarrow$ one bad step leads to many bad steps, restarting cures this
- In fact, restarts help a lot in proving convergence [3, p. 127], but almost a trick: the deflection "asymptotically vanish" and the gradient does all the work
- Typical restart after $n$ steps, not very nice when $n$ large (or small)
- Unrestarted P-R (not using $\beta^i_{PR+}$) does not converge for some $f$ [3, Th. 5.8]

## Conjugate gradient method: efficiency [3, p. 124]



- Powerful results for quadratic $f$:
  - superlinear convergence $a^{k+1} \leq [\,(\,\lambda_k - \lambda_n\,)\,/\,(\,\lambda_k + \lambda_n\,)\,]^{2k} a^1$ [3, Theorem 5.5]
  - only $k$ distinct eigenvalues $\Longrightarrow$ terminates in $k$ iterations [3, Theorem 5.4]
  - $k$ eigenvalues clustered around 1 $\Longrightarrow$ terminates in $n - k$ iterations [3, p. 116]
- For general $f$, efficiency $n$-step quadratic: $n$ CG steps $\approx$ 1 Newton step

  $\| x^{i+n} - x_* \| \leq r \| x^i - x_* \|^2$ [3, (5.51)]
- Makes sense: "close to $x_*$, $f \approx Q_{x_*}$" +

  "in $n$ steps the CG exactly solves a quadratic function"
- Not very nice when $n$ large
- Interesting relationships with quasi-Newton methods [3, §7.2], hybrid versions ...
- Variants surprisingly $\neq$ in practice; P-R/D-Y often better but varies a lot
- All in all: powerful approach, but not easy to manage

## Outline

- A "slightly" different process: $x^{i+1} \leftarrow x^i - \alpha^i \nabla f(x^i) + \beta^i (x^i - x^{i-1})$
  - $\beta^i (x^i - x^{i-1}) =$ "momentum term", keep $x^{i+1}$ going in same direction
  - while $-\nabla f(x^i)$ "force" steering the trajectory towards $x_*$ ($x^i$ "heavy")
- Large "momentum" $\beta^i \implies$ less "zig-zags" $\implies$ better convergence
- Hard to ensure $f(x^{i+1}) < f(x^i)$, in fact not a $f$-descent algorithm
- But with appropriate $\alpha^i$, $\beta^i$, a(n $\approx$) linear $d$-descent one:
  $d^{i+1} = \| x^{i+1} - x_* \| \approx\leq r \| x^i - x_* \| = d^i$ with $r = (\sqrt{\kappa} - 1) / (\sqrt{\kappa} + 1)$
- Optimal rate [1, Th. 3.15], can't do better (except "$\approx$", we'll see why)
- $\kappa = L / \tau = 1000 \implies$ this $r \approx 0.938$, gradient $r \approx 0.996$: may seem small, but $0.996^{100} = 0.6698$, $0.938^{100} = 0.0016$, and it can show in practice
- Geared towards $\alpha^i = \alpha$, $\beta^i = \beta$ constants
  (easy, inexpensive, but rigid: need to choose well)
- Requires specific (complicated) analysis, but main ideas seen already

- All starts from weird-ish two-terms recurrence definition of Heavy Ball:

$$\begin{bmatrix} x^{i+1} - x_* \\ x^i - x_* \end{bmatrix} = \begin{bmatrix} x^i + \beta^i ( x^i - x^{i-1} ) - \alpha^i ( \nabla f( x^i ) - \nabla f( x_* ) ) - x_* \\ x^i - x_* \end{bmatrix}$$

- Mean Value Theorem [5, Th. 5.4.5] applied to $\nabla f(\cdot) \implies \exists\, w^i \in [\, x_* \,,\, x^i \,]$ s.t $\nabla f( x^i ) - \nabla f( x_* ) = \nabla f^2( w^i )( x^i - x_* ) \implies$

$$\begin{bmatrix} x^{i+1} - x_* \\ x^i - x_* \end{bmatrix} = \begin{bmatrix} ( x^i - x_* ) - \alpha^i \nabla f^2( w^i )( x^i - x_* ) + \beta^i ( x^i - x^{i-1} ) \\ x^i - x_* \end{bmatrix} =$$

$$= \begin{bmatrix} [\, I - \alpha^i \nabla f^2( w^i ) ]( x^i - x_* ) + \beta^i ( x^i - x^{i-1} ) + \beta^i x_* - \beta^i x_* \\ x^i - x_* \end{bmatrix} =$$

$$= \begin{bmatrix} [\, I - \alpha^i \nabla f^2( w^i ) + \beta I \,]( x^i - x_* ) - \beta^i ( x^{i-1} - x_* ) \\ x^i - x_* \end{bmatrix} =$$

$$= \begin{bmatrix} ( 1 + \beta^i ) I - \alpha^i \nabla f^2( w^i ) & -\beta^i I \\ I & 0 \end{bmatrix} \begin{bmatrix} x^i - x_* \\ x^{i-1} - x_* \end{bmatrix}$$

- If we could find $\alpha^i$ , $\beta^i$ such that

$$\| \, C^i \, \| = \left\| \begin{bmatrix} ( 1 + \beta^i ) I - \alpha^i D^i & -\beta^i I \\ I & 0 \end{bmatrix} \right\| < 1 \;\;,\;\; D^i = \nabla f^2( w^i )$$

we would be done, but it's not that simple: $\| \, C^i \, \| > 1$

- $C^i$ not symmetric, $\|\, C^i \,\| \geq \rho(\, C^i\,)$ = spectral radius = $\max_j\{\, |\, \lambda_j(\, C^i\,)\,|\,\}$
  (careful: $\lambda_j(\, C^i\,)$ can be complex, $|\cdot|$ not the usual absolute value)

- $\rho(\, C^i\,) = \max_{j=1,\dots,n}\{\, \rho(\, C_j\,)\,\}$ with
  $$C_j = \begin{bmatrix} 1+\beta^i - \alpha^i\lambda_j(\, D\,) & -\beta^i \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$ (**check**) [nontrivial]

- Result: $\beta^i = \max\{\, |\, 1 - \sqrt{\alpha^i\tau}\,|\,,\, |\, 1 - \sqrt{\alpha^i L}\,|\,\}^2 \implies$ [**extremely** tedious]
  $\rho(\, C^i\,) \leq \sqrt{\beta^i} = \max\{\, |\, 1 - \sqrt{\alpha^i\tau}\,|\,,\, |\, 1 - \sqrt{\alpha^i L}\,|\,\}$ (**check**)

- $\alpha = 4\,/\,(\,\sqrt{L} + \sqrt{\tau}\,)^2 \implies \sqrt{\beta} = (\,\sqrt{L} - \sqrt{\tau}\,)\,/\,(\,\sqrt{L} + \sqrt{\tau}\,) < 1$
  $1\,/\,L \leq \alpha \leq 4\,/\,L$, growing as $L\,/\,\tau$ does, $\beta \leq 1$ (**check**)

- $r = \sqrt{\beta} = (\,\sqrt{\kappa} - 1\,)\,/\,(\,\sqrt{\kappa} + 1\,)$ optimal rate [1, Th. 3.15]

- This would be if we could prove linear convergence with $r = \sqrt{\beta}$,
  which is almost true but not quite

- Simplifying assumption: $f$ quadratic $\Longrightarrow$ $\nabla f$ costant $\Longrightarrow$ $C^i$ costant
$$\left\| \begin{bmatrix} x^{i+1} - x_* \\ x^i - x_* \end{bmatrix} \right\| \leq \| \, C^i \, \| \left\| \begin{bmatrix} x^1 - x_* \\ x^0 - x_* \end{bmatrix} \right\| \quad (i\text{-th power, by recursion})$$

- Gelfand's formula [6]: $\rho(\,C\,) = \lim_{i \to \infty} \| \, C^i \, \|^{1\,/\,i}$ (er ...eh?) $\Longrightarrow$
$\forall\, \varepsilon > 0 \;\exists h$ s.t. $\rho(\,C\,) - \varepsilon \leq \| \, C^i \, \|^{1\,/\,i} \leq \rho(\,C\,) + \varepsilon \;\; \forall\, i \geq h \Longrightarrow$
$\| \, C^i \, \| \leq (\, \rho(\,C\,) + \varepsilon\,)^i \Longrightarrow$ converges linearly if $\rho(\,C\,) + \varepsilon < 1$

- $\varepsilon$ arbitrary small provided $h$ "large": "sooner or later it starts converging" (but it may not at the beginning)

- The larger $h$, the more the convergence rate is closer to $\rho(\,C\,)$: quasi-linear convergence with rate $\rho(\,C\,)$

- Can be proven for general $L$-smooth $\tau$-convex $f$, we'll live to fight another day

- Works well in practice provided you find right $\alpha$ and $\beta$ (nontrivial)

- For non-convex $f$ converges if $\beta \in [\,0\,,\,1\,)$, $\alpha \in (\,0\,,\,2(1-\beta)\,/\,L\,)$ [4, p. 168] ($\beta$ "free" but $\alpha \to 0$ as $\beta \to 1$, and $2\,/\,L$ already rather small to start with)

- Can prove $O(\,1\,/\,i\,)$ error, not better than gradient (although better in practice with properly chosen $\alpha$, $\beta$)
- "Accelerated Gradient" has better theoretical convergence:

```
procedure y = ACCG ( f , x , ε )
  x₋ ← x; γ ← 1;
  do  {   // warning: black magic ahead
  γ₋ ← γ; γ ← ( √(4γ² + γ⁴) − γ² )/2; β ← γ( 1 / γ₋ − 1 );
  y ← x + β( x − x₋ ); g ← ∇f( y ); x₋ ← x; x ← y − (1 / L)g;
  } while( ‖ g ‖ > ε );
```

$\approx$ HB, except $\nabla f$ computed after momentum but before descent

- Optimal [1, Th. 3.14] $O(\,LD^2 \,/\, \sqrt{\varepsilon}\,)$ for $L$-smooth only [1, Th. 3.19], optimal linear $r = (\,\sqrt{\kappa} - 1\,)\,/\,(\,\sqrt{\kappa} + 1\,)$ if also $\tau$-convex [1, Th. 3.18]
- Non-monotone but can be made so (two $f$ computations per iteration)
- Complex theory, algorithm constructed to optimize worst-case behaviour
- In practice consistently slowish: carefully crafted to attain a given convergence speed, gets what it is constructed for

## Outline

- You can go (much) faster than gradient
- Thanks goodness, because gradient is very slow:
  convergence at best linear, possibly much worse if not $\tau$-convex
- There is only so much you can get with first-order methods
- Second-order methods have vastly better convergence ($\nearrow$ quadratic),
  but $\nabla^2 f$ has to $\exists$, be continuous, and you have to use it
- Although, you can use $\nabla^2 f$ without ever computing it
- First-and-a-half-order methods provide interesting trade-offs
- A lot of details need be considered, numerical aspects nontrivial
- Your mileage may vary

[1] S. Bubeck *Convex Optimization: Algorithms and Complexity*, arXiv:1405.4980v2, https://arxiv.org/abs/1405.4980, 2015

[2] S. Boyd, L. Vandenberghe *Convex Optimization*, https://web.stanford.edu/~boyd/cvxbook Cambridge University Press, 2008

[3] J. Nocedal, S.J. Wright, *Numerical Optimization – second edition*, Springer Series in Operations Research and Financial Engineering, 2006

[4] P. Ochs, *Local Convergence of the Heavy-Ball Method and iPiano for Non-convex Optimization* J. Optim. Theory Appl. 177, 153–180, 2018 https://doi.org/10.1007/s10957-018-1272-y

[5] W.F. Trench, *Introduction to Real Analysis* http://ramanujan.math.trinity.edu/wtrench/texts/TRENCH_REAL_ANALYSIS.PDF Free Hyperlinked Edition 2.04, December 2013

[6] Wikipedia – Gelfand's formula https://en.wikipedia.org/wiki/Spectral_radius#Gelfand's_formula

[7] Wikipedia – Rotation matrix
https://en.wikipedia.org/wiki/Rotation_matrix

[8] Wikipedia – Sherman-Morrison Formula
https://en.wikipedia.org/wiki/Sherman-Morrison_formula

## Outline

- $\varphi'_{x,d}(\alpha) = \langle d\,,\, \nabla f(x + \alpha d)\rangle$, hence $\varphi'_{x,d}(0) = \|d\|\|\nabla f(x)\|\cos(\theta)$. If $d$ where $-\nabla f(x)$ then $\theta = \pi$, since it's rotated by further 45 degrees ($\pi/2$), then either $\theta = 3\pi/4$ or $\theta = 5\pi/4$; in either case, $\cos(\theta) = -\sqrt{2}/2 =$ $= -\cos(\pi/4)$, hence $\varphi'_{x,d}(0) = -\|\nabla f(x)\|^2\cos(\pi/4)$ **[back]**

- $Q^i(d) = f(x^i) + \langle \nabla f(x^i)\,,\, d\rangle + \frac{1}{2}d^T\nabla^2 f(x^i)d$ , $\nabla Q^i(d^i) = 0 \;\equiv$ $\equiv\; \nabla f(x^i) + \nabla^2 f(x^i)d^i \;\equiv\; d^i = -[\nabla^2 f(x^i)]^{-1}\nabla f(x^i)$ **[back]**

- As in Theorem 1, $-\langle \nabla f(x^i)\,,\, d^i\rangle = (d^i)^T\nabla^2 f(x^i)d^i \geq \tau\|d^i\|^2$. By Taylor's theorem, $\lim_{d\to 0} R(d)/\|d\|^2 = 0 \;\equiv\; \forall\varepsilon > 0\;\exists\, h$ s.t. $R(d^i) \leq \varepsilon\|d^i\|^2$ $\forall i \geq h$. Thus, $R(d^i) \leq \varepsilon\|d^i\|^2 \leq (-\varepsilon/\tau)\langle \nabla f(x^i)\,,\, d^i\rangle$ Hence, $f(x^i + d^i) - f(x^i) = \frac{1}{2}\langle \nabla f(x^i)\,,\, d^i\rangle + R(d^i) \leq (\frac{1}{2} - \varepsilon/\tau)\langle \nabla f(x^i)\,,\, d^i\rangle =$ $(\frac{1}{2} - \varepsilon/\tau)\varphi'_{x^i,d^i}(0)$ eventually holds for all large enough $i$ however chosen $\varepsilon$. Thus, the Armijo condition $f(x^i + d^i) - f(x^i) \leq m_1\varphi'_{x^i,d^i}(0)$ will eventually hold at every iteration however chosen $m_1 < 1/2$
  This uses $\tau$-convexity, that is required for global convergence, but one could rather assume the milder $\nabla^2 f(x_*) \succ 0$ (required anyway for quadratic

convergence) and use $\lambda_n(\nabla^2 f(x_*)) > 0$ in place of $\tau$ at the cost of complicating the argument somewhat [**back**]

- $[H\sqrt{\Lambda}H^T][H\sqrt{\Lambda}H^T] = H\sqrt{\Lambda}[H^T H]\sqrt{\Lambda}H^T = H\Lambda H^T = Q$ [**back**]

- Obvious (we've seen it happening), but: $g = -z - R^{-1}q$, $z + g = -R^{-1}q$, $\nabla h(z+g) = (-R^{-1}q) + R^{-1}q = 0$ [**back**]

- ii) $\equiv$ $\nabla m^{i+1}(x) = \nabla f(x^{i+1}) + H^{i+1}(x - x^{i+1}) \implies$ $\nabla m^{i+1}(x^i) = \nabla f(x^i)$ $\equiv$ $\nabla f(x^{i+1}) + H^{i+1}(x^i - x^{i+1}) = \nabla f(x^i)$ $\equiv$ $\nabla f(x^{i+1}) - \nabla f(x^i) = H^{i+1}(x^{i+1} - x^i)$ $\equiv$ (S) [**back**]

- For any $f(\cdot)$, a finite difference approximation of the derivative $f'(\,x\,)$ can be computed as $(\,f(\,x+\varepsilon\,)-f(\,x\,)\,)\,/\,\varepsilon$ for some appropriately chosen "small" $\varepsilon$. Hence, this also holds for $\nabla^2 f(\cdot)$, which is the Jacobian of $\nabla f(\cdot)$. A finite difference approximation of the $i$-th column of $\nabla^2 f(\,x\,)$ can be computed as $(\,\nabla f(\,x+\varepsilon u^i\,)-\nabla f(\,x\,)\,)\,/\,\varepsilon$, $u^i$ as usual the $i$-the vector of the canonical basis; in other words, $[\,x+\varepsilon u^i\,]_h = x_h$ for all $h \neq i$, while $[\,x+\varepsilon u^i\,]_i = x_i+\varepsilon$. Computing this $H^0$ costs $n+1$ gradient computations (i.e., as many gradient computations as iterations, and $n$ can be large), plus it needs be inverted / factorised which is in general $O(\,n^3\,)$. Finding the appropriate numerical value for $\varepsilon$ is nontrivial either: too large and the approximation will be bad, too small and the numerical errors in the computation of $\nabla f(\cdot)$ will be so large that the noise overwhelms the signal and the approximation will be bad too [**back**]

- $\nabla f(\cdot) \in C^0$ and $d^{i-1}$ fixed $\Longrightarrow \lim_{\beta\to 0} \varphi'_{x^i, d^i(\,\beta\,)}(\,0\,) = \lim_{\beta\to 0}\langle\,\nabla f(\,x^i\,)\,,\,-\nabla f(\,x^i\,)+\beta^i d^{i-1}\,\rangle = -\|\,\nabla f(\,x^i\,)\,\|^2 < 0$ (otherwise the algorithm would have stopped already) [**back**]

- The first step is diagonalization of the upper-left block. $A = (1+\beta)I - \alpha D$ has eigenvalues $\lambda'_i = (1+\beta)I - \lambda_i$ and spectral decomposition $A = H\Lambda' H^T$ ($\lambda_i$, $H_i$ those of $D$); thus,

$$C' = \begin{bmatrix} H & 0 \\ 0 & H \end{bmatrix} \begin{bmatrix} A & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} H^T & 0 \\ 0 & H^T \end{bmatrix} = \begin{bmatrix} (1+\beta)I - \alpha\Lambda & -\beta I \\ I & 0 \end{bmatrix}$$

$H^T = H^{-1} \implies C'$ similar to $C \implies$ has the same eigenvalues [**?**]
Now, $C' \rightsquigarrow C''$ $2 \times 2$ block diagonal by exchanges of rows and columns

$$C'' = PC'P^T = \begin{bmatrix} C_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & C_n \end{bmatrix} \;,\; C_i = \begin{bmatrix} 1+\beta-\alpha\lambda_i & -\beta \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

$P$ permutation matrix $\implies P^T = P^{-1}$ [**?**] $\implies C''$ similar to $C' \implies$ eigenvalues of $C$ the union of those of $C_i$ [**back**]

- The eigenvalues of $C_i$ are the roots of the characteristic polynomial $p(\lambda) = \det(C_i - \lambda I) = \lambda^2 + (1 + \beta - \alpha\lambda_i)\lambda + \beta$. These are extremely tedious (but possible) to compute and write down, the use of a symbolic system is advised (see, e.g., the screenshot below). Once this is done, it is easy (with the symbolic system) to check that the largest of the two eigenvalues is always $\leq \sqrt{\beta}$ if $\beta \geq (1 - \sqrt{\alpha\lambda_i})^2$ (and $\leq$ something $> 1$, so wo don't care)

In[62]:= Eigenvalues$\left[\begin{pmatrix} 1+b-a & -b \\ 1 & 0 \end{pmatrix}\right]$

Out[62]= $\left\{\frac{1}{2}\left(1 - a - \sqrt{(-1+a-b)^2 - 4b} + b\right), \frac{1}{2}\left(1 - a + \sqrt{(-1+a-b)^2 - 4b} + b\right)\right\}$

In[60]:= Reduce$\left[\text{Abs}\left[\frac{1}{2}\left(1 - a - \sqrt{(-1+a-b)^2 - 4b} + b\right)\right] \leq \sqrt{b}\ \&\&\ a \geq 0\ \&\&\ b \geq 0, \{a, b\}\right]$

Out[60]= $(a == 0\ \&\&\ b == 1)\ ||\ \left(a > 0\ \&\&\ 1 - 2\sqrt{a} + a \leq b \leq 1 + 2\sqrt{a} + a\right)\ ||$
$\left(0 \leq a \leq 1\ \&\&\ \left(0 \leq b < 1 - 2\sqrt{a} + a\ ||\ b > 1 + 2\sqrt{a} + a\right)\right)\ ||$
$\left(a > 1\ \&\&\ b > 1 + 2\sqrt{a} + a\right)\ ||\ (a == 1\ \&\&\ b == 0)\ ||\ (0 \leq a < 1\ \&\&\ b == 0)$

In[61]:= Reduce$\left[\text{Abs}\left[\frac{1}{2}\left(1 - a + \sqrt{(-1+a-b)^2 - 4b} + b\right)\right] \leq \sqrt{b}\ \&\&\ a \geq 0\ \&\&\ b \geq 0, \{a, b\}\right]$

Out[61]= $(a == 0\ \&\&\ b == 1)\ ||\ \left(a > 0\ \&\&\ 1 - 2\sqrt{a} + a \leq b \leq 1 + 2\sqrt{a} + a\right)\ ||$
$\left(a > 1\ \&\&\ \left(0 \leq b < 1 - 2\sqrt{a} + a\ ||\ b == 0\right)\right)\ ||\ (a == 1\ \&\&\ b == 0)$

Hence, $\beta = \max_{i=1,\ldots,n}\{ ( 1 - \sqrt{\alpha\lambda_i} )^2 \} \implies$
$\rho( C ) \leq \sqrt{\beta} = \max\{ | 1 - \sqrt{\alpha^i\tau} | , | 1 - \sqrt{\alpha^i L} | \}$ **[back]**

- Since $0 < \tau \leq L$, $\alpha = 4 / ( \sqrt{L} + \sqrt{\tau} )^2 \leq 4 / ( \sqrt{L} )^2 = 4 / L$. On the other direction, $\alpha = 4 / ( \sqrt{L} + \sqrt{\tau} )^2 \geq 4 / ( \sqrt{L} + \sqrt{L} )^2 = 4 / ( 2\sqrt{L} )^2 = L$. Note that $\tau \to 0$ (very elongated level sets) $\implies \alpha \to 4 / L$ , while $\tau = L$ (perfectly circular level sets) $\implies \alpha = 1 / L$: the step is longer the more elongated are the level sets. For the rest, $\sqrt{\beta} = \max\{ | 1 - \sqrt{\alpha\tau} | , | 1 - \sqrt{\alpha L} | \} =$
$$= \max\left\{ \left| 1 - \sqrt{4\tau / ( \sqrt{L} + \sqrt{\tau} )^2} \right| , \left| 1 - \sqrt{4L / ( \sqrt{L} + \sqrt{\tau} )^2} \right| \right\} =$$
$$= \max\{ | ( \sqrt{L} + \sqrt{\tau} - 2\sqrt{\tau} ) / ( \sqrt{L} + \sqrt{\tau} ) | ,$$
$$| ( \sqrt{L} + \sqrt{\tau} - 2\sqrt{L} ) / ( \sqrt{L} + \sqrt{\tau} ) | \} =$$
$$= \max\{ | ( \sqrt{L} - \sqrt{\tau} ) / ( \sqrt{L} + \sqrt{\tau} ) | , | ( \sqrt{\tau} - \sqrt{L} ) / ( \sqrt{L} + \sqrt{\tau} ) | \} =$$
$= ( \sqrt{L} - \sqrt{\tau} ) / ( \sqrt{L} + \sqrt{\tau} ) \leq 1 \implies \beta \leq 1$ **[back]**