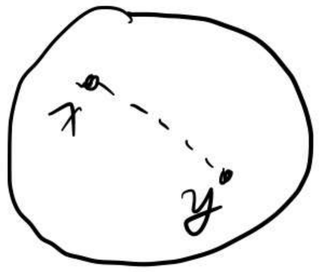
§1. Basics of Gradient Descent

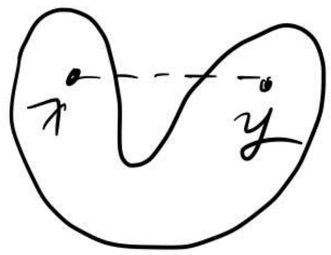
1. Def. A set $C \subseteq \mathbb{R}^d$ is said to be convex if $(1-\theta)x + \theta y \in C$. $\forall x, y \in C$, $\theta \in [0,1]$.

A function $f: \mathrm{dom}(f) \subseteq \mathbb{R}^d \to \mathbb{R}$ is said to be convex if $\mathrm{dom}(f)$ is convex and

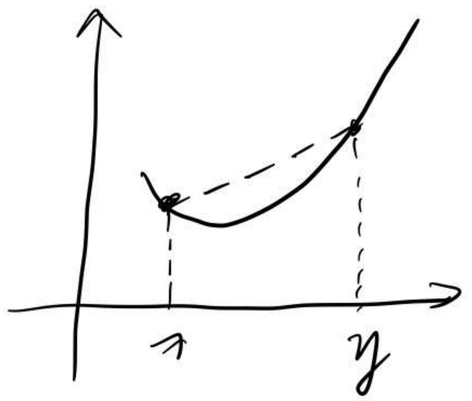
$$f((1-\theta)x + \theta y) \le (1-\theta) f(x) + \theta f(y). \quad \forall x, y \in \mathrm{dom}(f), \ \theta \in [0,1].$$

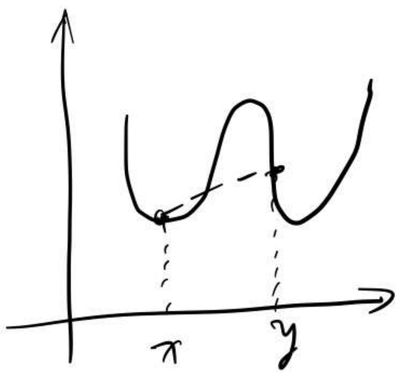


convex



non-convex





2. Lemma (Equivalent defs of convexity).

The following statements are equivalent.

(Suppose $f: \mathbb{R}^d \to \mathbb{R}$ is C^2).

(a) f is convex (b) Monotone gradient: $\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge 0 \quad \forall x, y \in \mathbb{R}^d$

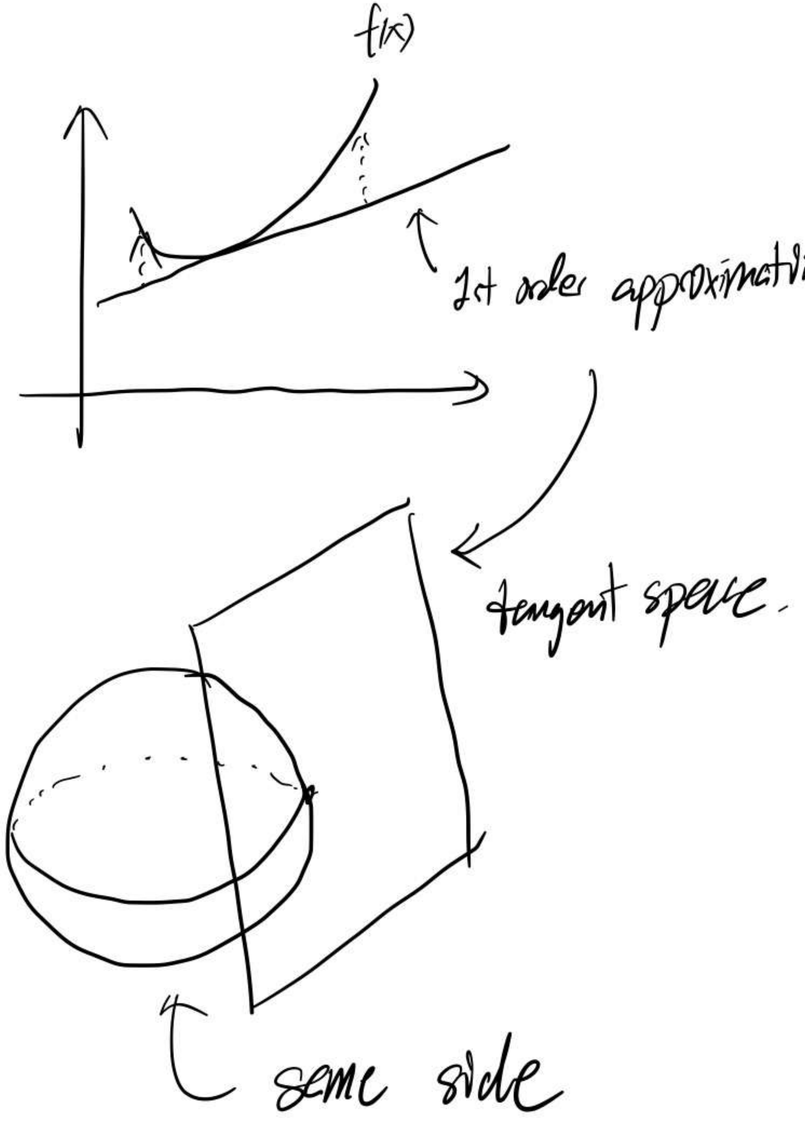
(c) Lower linear bound: $f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle \quad \forall x, y \in \mathbb{R}^d$.

(d) Positive curvature: $\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathbb{R}^d$.

Remarks.

(c) gradient $\Leftrightarrow$ local linear approximation

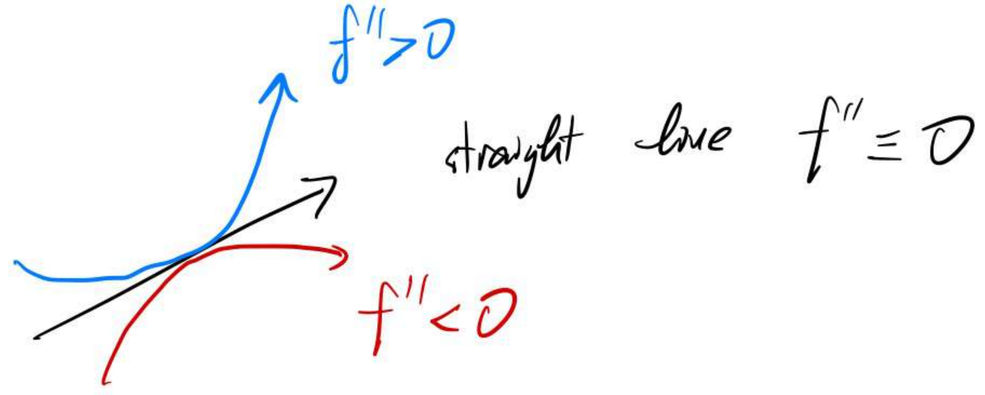
$\Leftrightarrow$ tangent space



(d) Hessian ($\Leftrightarrow$ local 2nd order approx.)

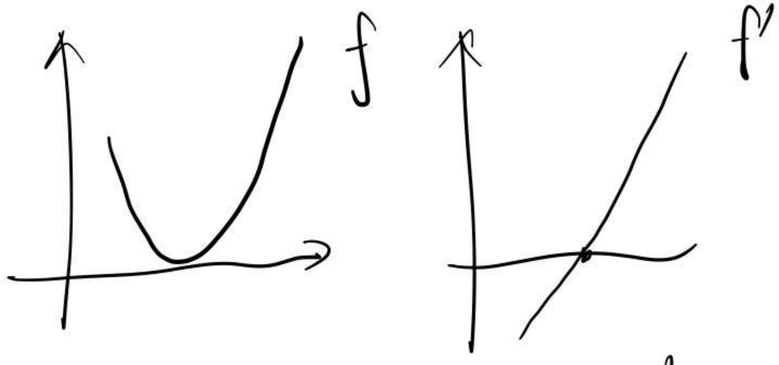
$\Leftrightarrow$ curvature

$\Leftrightarrow$ How "fast" it deviates from the tangent



Intuition: for a set to be convex.
the boundary should always
bend in the same way

(5) Monotone gradient.

1D 

"clear" that f convex $\Leftrightarrow$ f' monotone $\Leftrightarrow$ $(f'(x+\delta) - f'(x))\delta \geq 0 \quad \forall x, \delta \in \mathbb{R}$

∇f is monotone

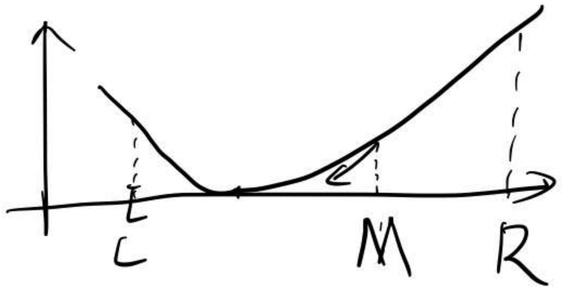
$\Leftrightarrow$ δx and $\delta(\nabla f)$ are positively correlated.

$\Leftrightarrow$ $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$

3. Ellipsoid method.

Q: How to minimize an 1D convex function?

A: Bisection!



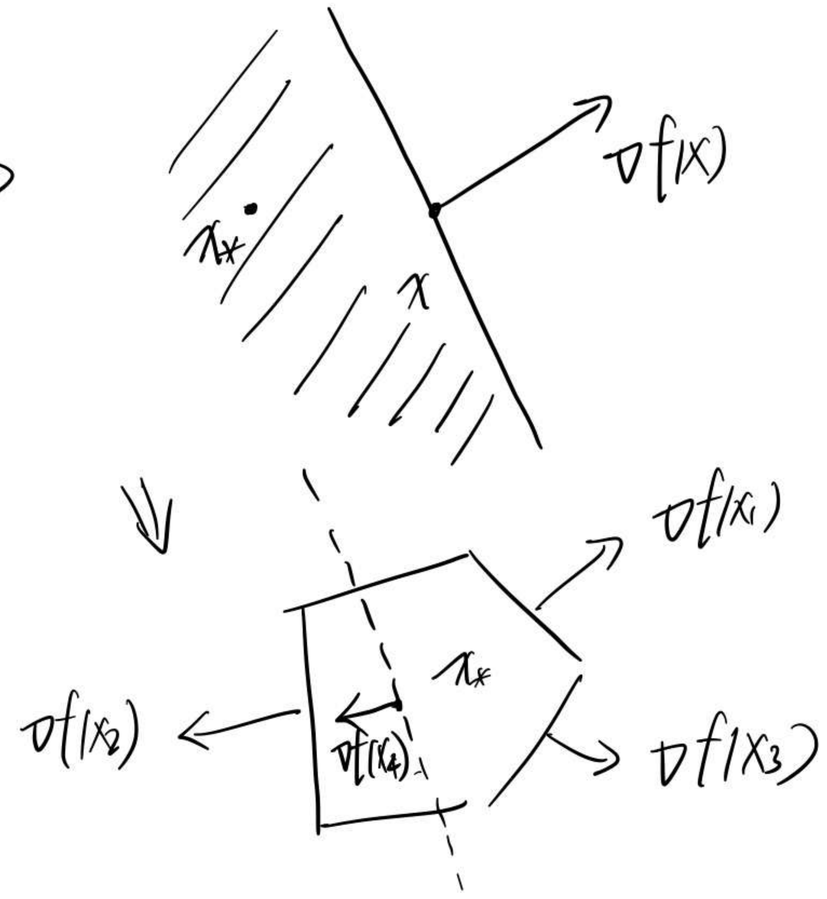
Q: "Bisection" on high dimension?

A: Ellipsoid method!

(Recall the monotonicity)

$\Rightarrow \langle \nabla f(x) - \nabla f(x_*), x - x_* \rangle \geq 0$

($\nabla f(x_*) = 0$)



Problem: It's hard to find a good point x_{k+1} that can halves the volume of the polytope.

Solution: Use ellipsoids instead of polytopes. and choose the center as the next point.

Claim: It's possible to construct E_k s.t.

$$\mathrm{Vol}(E_{k+1}) \leq \left(1 - \Omega\left(\frac{1}{d^2}\right)\right) \mathrm{Vol}(E_k)$$

4. Def/Lemma. (strong convexity).

The following statements are equivalent.

(a) f is μ-strongly convex. That is,

$x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex. ($\mu > 0$).

(b) $f((1-\theta)x + \theta y) \le (1-\theta)f(x) + \theta f(y) - \frac{\theta(1-\theta)\mu}{2}\|x-y\|^2$

(c) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \mu \|x-y\|^2$

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

(e) $\nabla^2 f(x) \succeq \mu I_d$.

Examples.

1) Linear functions are NOT strongly convex

2) Quadratic function $\frac{1}{2}x^T A x$ is $\lambda_{\min}(A)$ - strongly convex.

5. Def. (Gradient flow). Given a C^1 function $f: \mathbb{R}^d \to \mathbb{R}$, we define the gradient flow (GF) w.r.t. f with initial point $\vec{x} \in \mathbb{R}^d$ to be the solution x_t to the initial value problem:

$$\frac{d}{dt} x_t = -\nabla f(x_t), \quad x_0 = \vec{x}.$$

Remarks. (a) GF $\Leftrightarrow$ GD with infinitesimal step size

(b) By the Chain Rule,

$$\frac{d}{dt} f(x_t) = \langle \nabla f(x_t), \dot{x}_t \rangle = -\|\nabla f(x_t)\|^2 \leq 0$$

(GF never increases the function value.)

6. Lemma (Gronwall) Suppose that $\dot{u}_t \leq \alpha_t u_t$.

Then we have $u_t \leq u_0 \exp\left(\int_0^T \alpha_t \, dt\right)$

Linear system.

$$\dot{u}_t = \alpha_t u_t \Rightarrow u_T = u_0 \exp\left(\int_0^T \alpha_t \, dt\right)$$

$$\left(\dot{u}_t = A u_t \Rightarrow u_T = u_0 \exp(AT)\right)$$

Important: linear system

$\Rightarrow$ exponential growth/decrease.

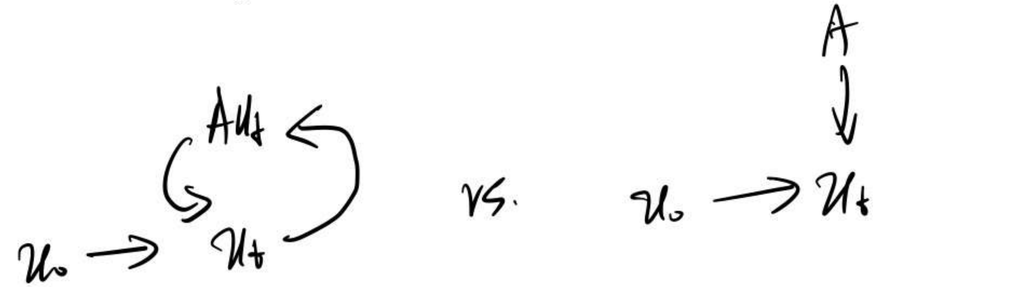
Example. (Discrete linear system)

$$u_{t+1} - u_t = 0.5 \, u_t$$

$$\Rightarrow u_{t+1} = 1.5 \, u_t = u_0 (1.5)^t$$

Compare linear systems with $\dot{u}_t = A \Rightarrow u_t = u_0 + AT$

$\Rightarrow$ linear growth/decrease.



7. Lemma Let $f:\mathbb{R}^d\to\mathbb{R}$ C^1, μ-strongly convex.
Let $x_* = \arg\min_x f(x)$. For any $\varepsilon>0$, we have
$\|x_T - x_*\| \le \varepsilon$ for all $T \ge \mu^{-1}\log\left(\frac{\|x_0-x_*\|}{\varepsilon}\right)$.

Pf. $\frac{d}{dt}\|x_t - x_*\|^2 = 2\langle x_t - x_*, \frac{d}{dt}x_t\rangle$

$= -2\langle x_t - x_*, \nabla f(x_t)\rangle$

$= -2\langle x_t - x_*, \nabla f(x_t) - \nabla f(x_*)\rangle$ ($\nabla f(x_*)=0$)

(strong convexity) $\le -2\mu\|x_t - x_*\|^2$

Gronwall's lemma $\Rightarrow \|x_T - x_*\|^2 = \|x_0 - x_*\|^2\exp(-2\mu T)$

□

Remark. strong convexity $\Rightarrow$ linear convergence rate

(optional)

8. Lemma (w/o. strong convexity) Let $f:\mathbb{R}^d\to\mathbb{R}$ C^1, convex. We have $f(x_T) \le f(x_*) + \frac{\|x_0-x_*\|^2}{2T}$.

Remark. When f is almost flat, the movement of x_t is slow. $\Rightarrow$ no linear contraction.
However, f convex $\Rightarrow$ $f(x_0)$ close to $f(x_*)$ in those regions.
$\Rightarrow$ track $f(x_t)$ directly.

Pf. Lower linear bound $\Rightarrow \langle\nabla f(x), x - x_*\rangle \ge f(x) - f(x_*)$

$\Rightarrow \frac{d}{dt}\|x_t - x_*\|^2 = -2\langle x_t - x_*, \nabla f(x_t)\rangle \le -2(f(x_t) - f(x_*))$

Integrate both sides

$\|x_T - x_*\|^2 - \|x_0 - x_*\|^2 \le -2\left(\int_0^T f(x_t) - Tf(x_*)\right)$

$\Rightarrow \frac{1}{T}\int_0^T f(x_t) \le f(x_*) + \frac{\|x_0-x_*\|^2}{2T}$

$\Rightarrow \ge f(x_T)$

□

(Optional)

- Lyapunov's direct method

Suppose that $\dot{x}_t = G(x_t)$ for some $G: \mathbb{R}^d \to \mathbb{R}^d$.

with $G(0) = 0$. If $\exists\, V: \mathbb{R}^d \to \mathbb{R}$ with

1) $V \in C^0$. 2) $V \geq 0$ with equality iff $x = 0$

3) $\langle G(x), V(x) \rangle \leq 0$ with equality iff $x = 0$

then $x_t \to 0$.

9. Polyak-Łojasiewicz (PL) inequality

Recall $\frac{d}{dt} f(x_t) = -\|\nabla f(x_t)\|^2$

Def. Let $f: \mathbb{R}^d \to \mathbb{R}$ be a C^1 function (not necessarily convex) with infimum f_*.
We say f satisfies the PL condition with PL constant $\mu > 0$ if

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu (f(x) - f_*)$$

($f(x) - f_*$: suboptimality)

Remark. PL $\Rightarrow$ linear convergence

Message. It suffices to lower bound $\|\nabla f(x_t)\|$

Another strategy. Constructing a descent direction u_t.

$$\|\nabla f(x_t)\| \geq \langle \nabla f(x_t), u_t/\|u_t\| \rangle$$

10. Lemma. μ-strongly convex $\Rightarrow$ μ-PL

Pf. μ-strong convexity

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

$$\Rightarrow \min_{y_1} f(y_1) \geq \min_{y_2} \left\{ f(x) + \langle \nabla f(x), y_2 - x \rangle + \frac{\mu}{2}\|y_2 - x\|^2 \right\}$$

LHS $= f_*$.

RHS $= f(x) + \frac{1}{2\mu}\|\nabla f(x)\|^2$

$$\Rightarrow \frac{1}{2}\|\nabla f(x)\|^2 \geq \mu (f(x) - f_*)$$ ☒

10'. Lemma. L-smoothness $\Rightarrow$ "reverse PL"

$$\frac{1}{2}\|\nabla f(x)\|^2 \leq L (f(x) - f_*)$$

(Present this after defining L-smoothness)

11. Def. Given C^2 $f: \mathbb{R}^d \to \mathbb{R}$, we define the gradient descent (GD) sterates of f with initial point $\vec{x}$ and step size $\eta > 0$ to be the sequence $(x_k)_{k \geq 0}$.

$$x_{k+1} = x_k - \eta \nabla f(x_k), \quad x_0 = \vec{x}.$$

GD vs GF. GD: $x_{k+1} = x_k - \int_0^\eta \nabla f(x_k)\, ds$

GF: $x_{(k+1)\eta} = x_{k\eta} - \int_0^\eta \nabla f(x_{k\eta+s})\, ds$

Observation. If ∇f doesn't change too fast ($\Rightarrow \nabla f(x_k) \approx \nabla f(x_{k\eta+s})$)

then GD $\approx$ GF.

12. Def. A C^2 function $f: \mathbb{R}^d \to \mathbb{R}$ is said to be L-smooth if its gradient is L-Lipschitz.

that is, $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$.

13. Lemma. The following statements are equivalent.

(a) f is L-smooth

(b) $\|\nabla^2 f(x)\|_2 \leq L$

(c) (Two-sided) upper quadratic bound.

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

4. Descent Lemma. $f:\mathbb{R}^d \to \mathbb{R}$, C^2. 2-smooth (not necessarily convex). Suppose $\eta \le 1/L$. We have

$$f(x_{k+1}) \le f(x_k) - \frac{\eta}{2}\|\nabla f(x_k)\|^2$$

Pf. Upper quadratic bound.

$$\Rightarrow f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), \underbrace{x_{k+1} - x_k}_{= -\eta \nabla f(x_k)} \rangle + \frac{L}{2}\|x_k - x_{k+1}\|^2$$

$$= f(x_k) - \eta\|\nabla f(x_k)\|^2 + \frac{L}{2}\eta^2\|\nabla f(x_k)\|^2$$

$$= f(x_k) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(x)\|^2$$

$$\le f(x_k) - \frac{\eta}{2}\|\nabla f(x_k)\|^2$$

∎

Remarks.

a) We only need $\eta < 2/L$ to guarantee $f(x_k) \downarrow$

b) This lemma can be vacuous in non-convex optimization

HW. loss: $f(w) = \frac{1}{2}(1 - w_1^2 w_2)^2$

Run GD with $\eta \in \{2/10, 2/12, 2/15\}$ and initial point $(0.7, 2)$ for 50 steps. Plot the loss value and sharpness along the GD trajectories.

Corollary. Within $\frac{2(f(x_0) - f(x_T))}{\eta\varepsilon}$ iterations, GD with $\eta \le 1/L$ can find a point with $\|\nabla f(x_k)\|^2 \le \varepsilon$.

Pf. Sum both sides and rearrange terms.

Remark. This corollary, combined with strong convexity $\Rightarrow$ PL, gives a convergence result for GD, though the rate obtained this way is not optimal.

Remark. For well-conditioned f, this matches the rate of GF.

16. Lemma. Let $f: \mathbb{R}^d \to \mathbb{R}$, C^1, μ-strongly convex, and L-smooth. Choose $\eta \le 1/L$. We have

$$\|x_{k+1} - x_*\|^2 \le (1-\eta\mu)^k \|x_0 - x_*\|^2.$$

(Compare this lemma and its proof to the GF ver.)

Pf. $\|x_{k+1} - x_*\|^2 = \|x_k - x_* - \eta \nabla f(x_k)\|^2$

$$= \|x_k - x_*\|^2 - 2\eta \langle x_k - x_*, \nabla f(x_k)\rangle + \frac{\eta^2}{2}\|\nabla f(x_k)\|^2$$

(Lower bound) $\ge f(x_k) - f_* + \frac{\mu}{2}\|x_k - x_*\|^2$ (reverse PL) $\le 2L(f(x_k) - f_*)$

discretization error

$$\le (1-\eta\mu)\|x_k - x_*\|^2 - 2\eta(1-\eta L)(f(x_{(k)}) - f_*) \ge 0$$ ∎