LECTURE NOTES ON OPTIMIZATION BASICS [*]

A. A. ERGÜR, A. HOOKER, J.T. ANDERSON

Abstract. This scribe contains lecture notes on optimizing and proving convexity for functions via methods like gradient decent, Taylor Expansion, First-Order Necessary Condition for optimality, ect.

Key words. fill in the keywords here

AMS subject classifications. safely ignore

1. Notes.

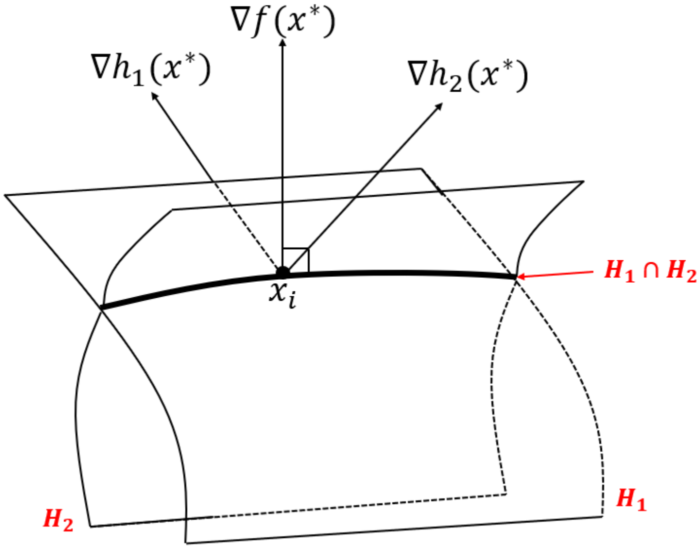

FIG. 1. *Illustration of constrained optimization geometry. The black curve represents the intersection $H_1 \cap H_2$, where $H_1 = \{x \in \mathbb{R}^n : h_1(x) = 0\}$ and $H_2 = \{x \in \mathbb{R}^n : h_2(x) = 0\}$ are constraint surfaces. At the point $x_i \in H_1 \cap H_2$, the gradients $\nabla h_1(x^*)$ and $\nabla h_2(x^*)$ are normal to their respective constraint surfaces. The gradient of the objective function $\nabla f(x^*)$ lies in the span of the constraint gradients, enforcing the condition $\nabla f(x^*) = \lambda_1 \nabla h_1(x^*) + \lambda_2 \nabla h_2(x^*)$, as per the method of Lagrange multipliers.*

$$c : \mathbb{R} \to H_1 \qquad h_1(c(t)) = g_1(t) = 0 \qquad g_1'(t) = 0$$

Let constraint surfaces H_1 and H_2 be defined by:

$$H_1 := \{x \in \mathbb{R}^n : h_1(x) = 0\}, \qquad H_2 := \{x \in \mathbb{R}^n : h_2(x) = 0\}$$

$$\text{e.g. } \{x \in \mathbb{R}^n, \quad x_1^2 + x_2^2 - 2x_1x_2 - 5 = 0\}$$

[*] We thank Robbins family for supporting the Algorithmic Foundations of Data Science Course

Let $c : \mathbb{R} \to H_1 \subset \mathbb{R}^n$ be a curve on H_1. Then:

$$h_1(c(t)) = g_1(t) = 0 \quad \Rightarrow \quad g_1'(t) = 0$$

Example $\{x \in \mathbb{R}^n, \quad x_1^2 + x_2^2 - 2x_1x_2 - 5 = 0\}$
Let c be a curve on H_1:

$$c : \mathbb{R} \to \mathbb{R}^n, \quad \text{and} \quad h_1(c(t)) = 0$$

$$g_1 : \mathbb{R} \to \mathbb{R}, \quad g_1(t) = h_1(c(t))$$

$$g_1'(t) = ?$$

Directional Derivative.

$$D_v h_1(x) = \lim_{\delta \to 0} \frac{h_1(x + \delta v) - h_1(x)}{\delta}$$

$$D_v h_1(x) = \langle \nabla h_1(x), v \rangle$$

Taylor Expansion at x_0.

$$h_1(x) \approx h_1(x_0) + \langle \nabla h_1(x_0), x - x_0 \rangle + \frac{1}{2}(x - x_0)^T \nabla^2 h_1(x)(x - x_0)$$

Curve Approximation

$$c(t + \delta) = c(t) + \delta \cdot c'(t) + O(\delta^2)$$

$\blacksquare$

Differentiating Along the Curve.

$$\frac{d}{dt} h_1(c(t)) = D_{c'(t)} h_1(c(t)) = 0$$

$$0 = \langle \nabla h_1(c(t)), c'(t) \rangle$$

Tangent Space.

$$H_1 = \{x \in \mathbb{R}^n : h_1(x) = 0\}$$

$$T_x H_1 := \{v \in \mathbb{R}^n : \langle \nabla h_1(x), v \rangle = 0\}$$

$$T_x(H_1 \cap H_2) := \{v \in \mathbb{R}^n : \langle \nabla h_1(x), v \rangle = 0,\ \langle \nabla h_2(x), v \rangle = 0\}$$

$$\langle v, \nabla h_1(x) + 2\nabla h_2(x) \rangle = 0$$

$$\langle \lambda_1 \nabla h_1(x) + \lambda_2 \nabla h_2(x), v \rangle = 0$$

$$\min_{x \in H_1 \cap H_2} f(x) = f(x^*)$$

$$Df(x^*) = \lambda_1 \nabla h_1(x^*) + \lambda_2 \nabla h_2(x^*)$$

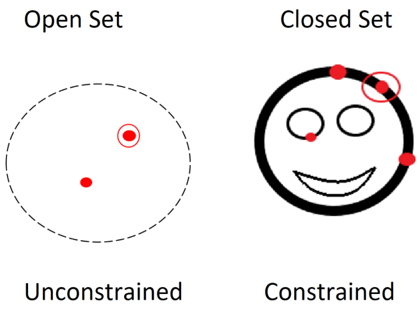

FIG. 2. *Illustration of open vs. closed sets in optimization. The left diagram represents an unconstrained optimization problem over an open set, where local minima must lie strictly inside the domain. The right diagram shows a constrained optimization problem over a closed set, where local minima may lie on the boundary. In closed sets, feasible regions include boundary points, enabling constrained minima that are not accessible in open domains.*

Defining Optimization Problem. Consider the difference between an open set (unconstrained) and a closed set (constrained).

If for all x such that $\|x - x^*\| < \varepsilon$, we have:

$$f(x^*) \leq f(x)$$

$$f(x^*) \rightarrow \text{local minimum}$$

Strict Local Minimum. If

$$f(x^*) < f(x)$$

then

$$f(x^*) \rightarrow \text{strict local minimum}$$

First-Order Necessary Condition for Optimality. If $f(x)$ has a local minimum at $x^* \in H$, where

$$H := \{x \in \mathbb{R}^n : h_1(x) = 0,\ h_2(x) = 0\}$$

then

$$Df(x^*) = \lambda_1 \nabla h_1(x^*) + \lambda_2 \nabla h_2(x^*)$$

for some λ_1 and λ_2.

Example.

$$f(x_1, x_2) = x_1^2 + 2x_2 + 3, \qquad \nabla f(x) = (2x_1, 2)$$

$$H := \left\{x \in \mathbb{R}^2 : 3x_1 + 2x_2 = 5\right\}$$

$$h_1(x) = 3x_1 + 2x_2 - 5 = 0$$

$$\nabla h_1(x) = (3, 2)$$

$$\text{e.g., } (2x_1, 2) = \lambda(3, 2)$$

Second-Order Necessary Conditions. If $f(x)$ has a local minimum at $x^* \in H$, then for every $v \in T_{x^*}H$, we have:

$$v^T \nabla^2 f(x^*) v \geq 0$$

Why?. Using a second-order Taylor expansion:

$$f(x^* + v) \approx f(x^*) + \langle Df(x^*), v\rangle + v^T \nabla^2 f(x^*) v$$

If $f(x^* + tv) \approx f(x^*) + v^T \nabla^2 f(x^*) v$, then the second-order term dominates when $Df(x^*) = 0$.

Example.

$$H := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$$

$$f(x) = \langle Ax, x\rangle\,, \quad A \in \mathbb{R}^{n\times n}$$

Convexity.

Convex Set.

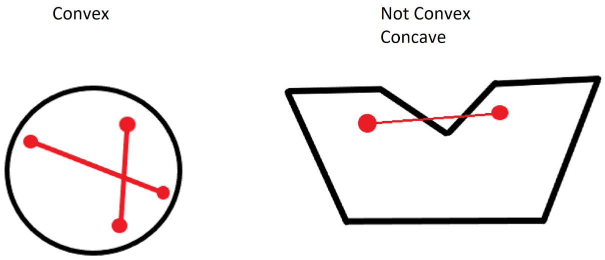

FIG. 3. *Comparison of convex and non-convex (concave) sets. On the left, a convex set is shown where any line segment connecting two points in the set lies entirely within the set. On the right, a non-convex (concave) set is illustrated where a line segment between two points exits the boundary, violating the convexity condition. Convexity plays a key role in optimization, as local minima in convex sets are also global minima.*

For every $x, y \in K$ and every $t \in (0, 1)$, we have:

$$\lambda x + (1 - \lambda) y \in K$$

Convex Function. Let $f : \mathbb{R}^n \to \mathbb{R}$, or more generally $f : K \to \mathbb{R}$, where $K \subseteq \mathbb{R}^n$ is convex.

Define the epigraph of f:

$$\operatorname{Epi} f := \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : t \geq f(x)\}$$

Then f is a convex function if $\operatorname{Epi} f$ is a convex set.

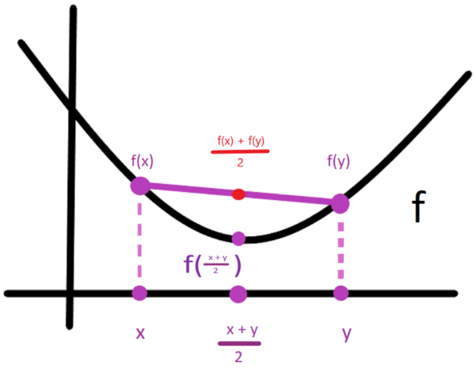

Fig. 4. *Convex function. For any two points x and y in the domain, the function satisfies $f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2}$. This inequality demonstrates that the graph of a convex function lies below the chord connecting $(x, f(x))$ and $(y, f(y))$. This is equivalent to the definition that the epigraph of f is a convex set.*

Equivalent Definition of Convexity. A function f is convex if for all $\lambda \in (0,1)$ and all x, y, we have:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

First-Order Characterization. f is convex on a convex domain K if and only if for all $x, y \in K$:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

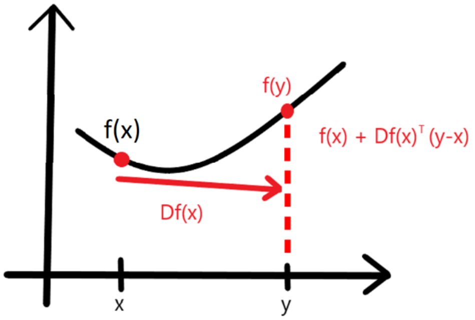

Fig. 5. *First-order condition for convexity. A function f is convex if its graph lies above all of its tangents. For all $x, y \in K$, the inequality $f(y) \geq f(x) + \nabla f(x)^T (y - x)$ holds. The figure shows the tangent line at x (in red), and demonstrates that the function value at y lies above this tangent, confirming convexity.*

Second-Order Characterization on Constraint Sets. Let

$$\Omega := \{x \in \mathbb{R}^n : h_1(x) \geq 0, \ldots, h_m(x) \geq 0\}$$

and let $f : \Omega \to \mathbb{R}$.

Then f is convex if and only if for all $x \in \Omega$ and all $v \in T_x\Omega$, we have:

$$v^T \nabla^2 f(x) v \geq 0$$

Strict Convexity. A function f is strictly convex if:

$$v^T \nabla^2 f(x) v > 0$$

for all $v \in T_x\Omega \setminus \{0\}$
Or equivalently:

$$\nabla^2 f(x)\big|_{T_x\Omega} \succeq m \cdot I$$

for some $m > 0$, where I is the identity matrix.

Convex Geometry (3). Let K be a domain.
Interior:.

$$\text{int } K := \{x \in K : \exists \varepsilon > 0 \text{ such that } B(x, \varepsilon) \subseteq K\}$$

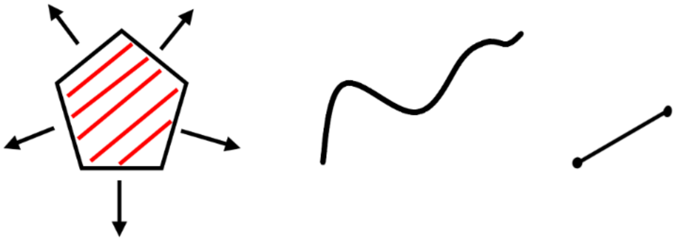

FIG. 6. *Geometric illustration of convex set properties. The figure depicts several key geometric concepts for a set K: the interior (int K), where small open balls are fully contained in K; the affine hull (aff(K)), the smallest affine space containing K; the convex hull (conv(K)), the smallest convex set containing all points in K; and the relative interior (relint(K)), which is the interior relative to the affine hull. These constructions are foundational in convex analysis and constrained optimization.*

Affine Hull:.

$$\text{affine hull of } K := \left\{\sum d_i x_i : \sum d_i = 1,\ x_i \in K\right\}$$

Convex Hull:.

$$\text{convex hull of } K := \left\{\sum t_i x_i : t_i \geq 0,\ \sum t_i = 1,\ x_i \in K\right\}$$

Relative Interior:.

$$\text{relint } K := \{x \in K : \exists \varepsilon > 0 \text{ such that } B(x, \varepsilon) \cap \text{aff}(K) \subseteq K\}$$

For $x \in \text{relint}(K)$, the effective constraints are the ones on aff(K).

Convex Optimization (4). Let Ω be a closed and convex domain. If f is convex on Ω, then any local solution to

$$\min_{x\in\Omega} f(x)$$

is also a global solution.

Moreover, the set

$$\left\{x \in \Omega : f(x) = \min_{x\in\Omega} f(x)\right\}$$

is a convex set.

(For concave f, the same claims hold but for $\max_{x\in\Omega} f(x)$.)

Why?.

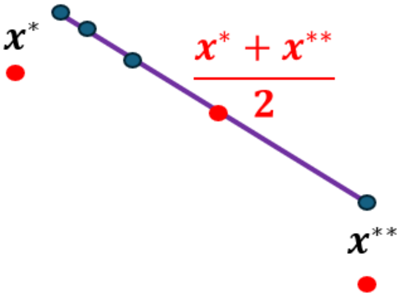

Fig. 7. *Global optimality of local minima for convex functions. For a convex function f defined over a convex set Ω, any local minimum x^* must also be a global minimum. The figure illustrates this by showing that for a local minimum x^* and a global minimum candidate x^{**}, the point halfway between them has a function value strictly less than $f(x^*)$, violating local minimality—unless $f(x^*) = f(x^{**})$.*

Let x^* be a local minimum, and suppose x^{**} is a global minimum. Then:

$$f\left(\frac{x^* + x^{**}}{2}\right) \leq \frac{1}{2}f(x^*) + \frac{1}{2}f(x^{**}) < f(x^*)$$

This contradicts the assumption that x^* is a local minimum unless $f(x^*) = f(x^{**})$, i.e., x^* is also a global minimum.

Consequence and Amendment. Consequence: If $x \in \mathbb{R}^n$, f is convex, and $f : \mathbb{R}^n \to \mathbb{R}$, then

$$\nabla f(x^*) = 0 \quad \Rightarrow \quad x^* \text{ is a global minimizer.}$$

Amendment: If f is strictly convex (or concave), then the minimizer (or maximizer) is *unique.*

Example. Let

$$\Omega := \{x \in \mathbb{R}^n : x_1 + \cdots + x_n \geq 1,\ x_i \geq 0 \text{ for all } i\}$$

Define:

$$H(x) = -\sum_{i=1}^{n} x_i \ln x_i, \qquad \max_{x\in\Omega} H(x) = ?$$

Gradient:

$$\nabla H(x) = -(\ln x_1 + 1, \ln x_2 + 1, \ldots, \ln x_n + 1)$$

Note: $(t \ln t)' = \ln t + t \cdot \frac{1}{t} = \ln t + 1$, so:

$$\frac{\partial}{\partial x_i}(-x_i \ln x_i) = -\ln x_i - 1$$

Hessian:

$$\nabla^2 H(x) = \begin{bmatrix} -\frac{1}{x_1} & 0 & \cdots & 0 \\ 0 & -\frac{1}{x_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{x_n} \end{bmatrix} < 0$$

Since H is strictly concave, this implies a unique maximizer.

$$a = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right), \qquad b = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$$

Necessary Conditions of Optimality for Convex Domains. Let Ω be a convex domain, and $f \in C^1$ (continuously differentiable). For any $x \in K$, we define the *normal cone*:

$$\mathcal{N}_\Omega(x) := \left\{ v \in \mathbb{R}^n : v^T(y - x) \leq 0 \quad \forall y \in \Omega \right\}$$

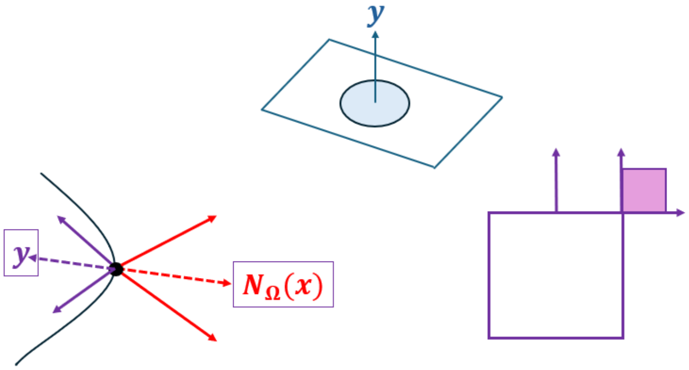

FIG. 8. *Geometric interpretation of the normal cone $\mathcal{N}_\Omega(x)$ to a convex set Ω at point x. The normal cone consists of all vectors $v \in \mathbb{R}^n$ such that $v^T(y-x) \leq 0$ for all $y \in \Omega$. In the illustrations, red arrows indicate valid normal directions at boundary points of the set, while dashed vectors show that directions outside Ω violate the condition. This concept underlies optimality conditions in constrained convex optimization.*

Then, if

$$x^* = \arg\min_{x\in\Omega} f(x)$$

we have the condition:

$$-\nabla f(x^*) \in \mathcal{N}_\Omega(x^*)$$

Example:. Q: Let

$$K := \left\{x \in \mathbb{R}^2 : 0 \le x_1 \le 1,\ 0 \le x_2 \le 1\right\}$$

Please describe the collection of linear functions

$$f(x) = \langle c, x\rangle$$

such that

$$(1,0) = \arg\min_{x\in K} f(x)$$

Descent Methods. A vector v is a **descent direction** for x if

$$f(x+tv) < f(x)$$

for all sufficiently small $t > 0$.

Lemma. If f is continuously differentiable around x, then any vector v such that

$$\langle \nabla f(x), v\rangle < 0$$

is a descent direction.

Question:. Among all v such that $\|v\|_2 = 1$, which one is the steepest descent direction at x?

$$\boxed{\frac{-\nabla f(x)}{\|\nabla f(x)\|}}$$

Gradient Descent (a.k.a. Steepest Descent). The iterative update rule is:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

Questions:
- How many steps to converge?
- What should the value of α_k (Learning Rate) be? (learning rate)

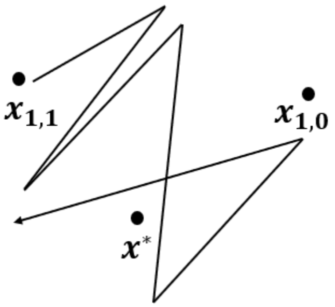

FIG. 9. *Illustration of zig-zagging behavior in optimization trajectories. The figure shows how an optimization path (e.g., gradient descent) may alternate directions inefficiently when approaching the solution x^*, particularly in cases with ill-conditioned level sets or suboptimal step directions. The points $x_{1,0}$ and $x_{1,1}$ represent successive iterates that deviate significantly due to sharp curvature or misaligned gradients.*

Definition: Lipschitz and Smooth Functions. Lipschitz Continuity:

$$\|f(x) - f(y)\| \leq L\|x - y\| \quad \Rightarrow \quad f \text{ is } L\text{-Lipschitz}$$

Smoothness:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \Rightarrow \quad f \text{ is } L\text{-smooth}$$

How to Know?.

- If on the line segment between x and y, we have
$$\|\nabla f(z)\| \leq L \quad \text{for all } z \text{ on the segment}$$
then f is L-Lipschitz.
- If on the line segment between x and y, we have
$$-LI \preceq \nabla^2 f(z) \preceq LI \quad \text{for all } z \text{ on the segment}$$
then f is L-smooth.

Lemma: Smoothness Inequality. If f is L-smooth, then:

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2$$

Why? Taylor's theorem.

Gradient Descent on an L-smooth Function. The update rule is:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

Now apply the smoothness inequality:

$$\begin{aligned} f(x_{k+1}) &= f(x_k) + \nabla f(x_k)^T(-\alpha_k \nabla f(x_k)) + \frac{L}{2}\| - \alpha_k \nabla f(x_k)\|^2 \\ &= f(x_k) - \alpha_k\|\nabla f(x_k)\|^2 + \frac{L}{2}\alpha_k^2\|\nabla f(x_k)\|^2 \end{aligned}$$

Choosing $\alpha_k = \frac{1}{L}$, we get:

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

This gives a descent guarantee.

Iterate $f(x)$ from $x_0 \to x_1 \to \cdots \to x_t$ to reduce function value.

Gradient Descent Convergence Bound. From the descent guarantee:

$$f(x_t) \leq f(x_0) - \frac{1}{2L}\sum_{i=0}^{t-1}\|\nabla f(x_i)\|^2$$

Then:

$$\sum_{i=0}^{t-1}\|\nabla f(x_i)\|^2 \leq (f(x_1) - f(x_0)) \cdot L$$

$$\sum_{i=0}^{t-1} \|\nabla f(x_i)\|^2 \leq (f(x_t) - f(x^*)) \cdot L$$

$$\boxed{\min_i \|\nabla f(x_i)\|^2 \leq \frac{L}{t}\left(f(x_t) - f(x^*)\right)}$$

Convergence for Convex and L-Smooth Functions. If f is convex and L-smooth, and we use gradient descent with

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

then we have:

$$f(x_t) - f(x^*) \leq \frac{L}{2t}\|x_0 - x^*\|^2$$

Why?. From the descent guarantee:

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

And from convexity:

$$f(x^*) \geq f(x_k) + (x^* - x_k) * \nabla f(x_k)$$

So: (*)

$$\boxed{f(x_{k+1}) \leq f(x^*) + (x_k - x^*)\nabla f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2}$$

Note:.

$$x_{k+1} - x^* = x_k - x^* - \frac{1}{L}\nabla f(x_k)$$

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 + \frac{1}{L^2}\|\nabla f(x_k)\|^2 - \frac{2}{L}\nabla f(x_k)(x_k - x^*)$$

(**)

$$\boxed{\nabla f(x_k)^T(x_k - x^*) - \frac{1}{2L}\|\nabla f(x_k)\|^2 = \frac{L}{2}\left(\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2\right)}$$

From Previous (*) and (**). ∎

$$\sum_{k=0}^{t-1}(f(x_{k+1}) - f(x^*)) \leq \frac{L}{2}\sum_{k=0}^{t-1}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)$$

$$\frac{1}{t}\sum_{k=0}^{t-1}(f(x_{k+1}) - f(x^*)) \leq \frac{L}{2t}\|x_t - x^*\|^2$$

$$\boxed{f(\bar{x}_t) \leq \frac{1}{t}\sum_{k=0}^{t-1} f(x_{k+1}) \leq f(x^*) + \frac{L}{2t}\|x_t - x^*\|^2}$$

Convexity.

$$t \geq \frac{L}{2\varepsilon}\|x_0 - x^*\|^2 \Rightarrow f(x_t) - f(x^*) \leq \varepsilon$$

Strong Convexity. Recall for strongly convex f, we have:

$$f(z) \geq f(y) + \nabla f(y)^T(z-y) + \frac{1}{2}m\|z-y\|^2$$

Then to ensure $f(x_t) - f(x^*) \leq \varepsilon$, it suffices to choose:

$$t \geq \frac{L}{m}\ln\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right)$$

$$\Rightarrow \quad f(x_t) - f(x^*) \leq \varepsilon$$

How to Write Down a Descent Algorithm?. We use the update rule:

$$x_{k+1} = x_k - \alpha_k v_k$$

Find $v_k \in \mathbb{R}^n$, $\alpha_k \in \mathbb{R}$ such that:

$$f(x_{k+1}) \leq f(x_k) - c \cdot \|\nabla f(x_k)\|^2 \quad \text{for some } c > 0$$

Finding v_k. We want:

$$0 < \varepsilon < -\frac{\nabla f(x_k)^T v_k}{\|\nabla f(x_k)\| \cdot \|v_k\|}$$

and

$$0 < \gamma_1 \leq \frac{\|v_k\|}{\|\nabla f(x_k)\|} \leq \gamma_2$$

How? Choosing v_k.

1. **Gauss–Southwell Rule:**

   $$v_k = -[\nabla f(x_k)]_{i_k} \cdot e_{i_k}$$

   where

   $$i_k := \arg\max_{1 \leq j \leq n} |\nabla f(x_k)_j|$$

   In practice, useful when updates are of specific type.
2. **Randomized v_k:**
   Lazy approach:
   - Pick $i \in \{1, 2, \ldots, n\}$ randomly
   - Set $v_k := \nabla f(x_k)_i$

   Then:

   $$\mathbb{E}\left[\nabla f(x_k)^T v_k\right] \geq \frac{1}{n}\|\nabla f(x_k)\| \cdot \|v_k\|$$
3. **Another Approach:** Find a ”cheap” stochastic estimator $g(x, \xi)$ such that ξ_k is a random variable with:

   $$\mathbb{E}\left[g(x_k, \xi_k)\right] = \nabla f(x_k)$$

   Then set:

   $$v_k := -g(x_k, \xi_k)$$

Finding Step Length α_k.

1. **Fixed Step Length**
2. **Exact Line Search**

$$\min_{t>0} f(x_k + tv_k)$$

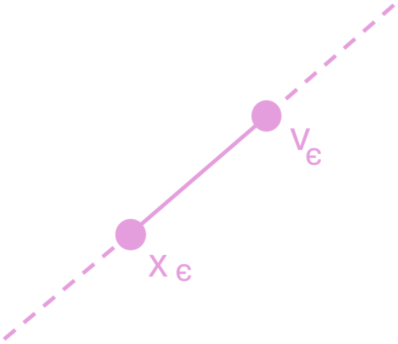

FIG. 10. *Visualization of exact line search. Starting from the current iterate x_ε, we search along the descent direction v_ε to minimize the objective function $f(x + tv)$. The goal is to find the optimal step size t that minimizes f along the line defined by $x + tv$. This approach ensures the most progress in the given direction and forms the basis of several optimization algorithms.*

3. **Approximate Line Search** Use Wolfe conditions:
   - **(1) Sufficient decrease (Armijo condition):**

$$f(x_k + \alpha v_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^T v_k$$

   - **(2) Curvature condition:**

$$\nabla f(x_k + \alpha v_k)^T v_k \geq c_2 \nabla f(x_k)^T v_k$$

   where $0 < c_1 < c_2 < 1$
4. **Backtracking Line Search**
   - Start with an initial guess $\tilde{\alpha} > 0$
   - Choose a parameter $\beta \in (0, 1)$
   - Check:

$$f(x_k + \tilde{\alpha} v_k), \quad f(x_k + \tilde{\alpha}\beta v_k), \quad \ldots, \quad f(x_k + \tilde{\alpha}\beta^t v_k)$$

   until the Armijo condition (1) is satisfied.

Note: Many algorithms based on this framework have been developed since the 1970s.

2. Exercises.

2.1. Question 1.

For a diferentiable function $f : \mathbb{R}^n \to \mathbb{R}, x \in \mathbb{R}^n$ find the vector v such that $||v||_2 = 1 and D_v f(x) \geq D_u f(x)$ for all $u \in \mathbb{R}^n, ||u||_2 = 1$

2.2. Question 2. $f(x) = \langle x, Ax \rangle$. We want to minimize f in the sphere

$$S^{n-1} := \{x \in \mathbb{R}^n : ||x||_2 = 1\}$$

. Write down first and second order necessary conditions for obtaining a local minimum of f at a vector $x \in S^{n-1}$

2.3. Question 3. $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function and Find a vector v such that

$$||v||_2 = 1$$

, and

$$f(x) + \langle v, \nabla f(x) \rangle + 1/2v^T \nabla^2 f(x)v \leq f(x) + \langle u, \nabla f(x) \rangle + 1/2u^T \nabla^2 f(x)u$$

for all $u \in \mathbb{R}^n, ||u||_2 = 1$

2.4. Question 4. Revisit the entropy maximimization problem in class. Solve it yourself.

2.5. Question 5. $K := \{x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n : -1 \leq x_i \leq 1\}$ is a cube. We know that a differentiable function f has a local minima on K at $(-1, -1, -1, ..., -1)$. What can you say about f? Characterize all functions that can have a local min at $-1, -1, -1, ..., -1$.

2.6. Question 6. This exercise is meant to make you revisit analysis of gradient descent for a nice convex and L-smooth function f. The analysis is included for example in the lecture notes below, and the Recht-Wright textbook below as well.

We define $x_{k+1} = x_k - 1/L\nabla f(x_k)$ Assume that we have for initial starting point x_0, and the optimal point x^* we have $||x_0 - x^*||_2 \leq 1$

We want to find a point x_i such that $||x_i - x||_2 \leq \epsilon$. How many iterations suffice?