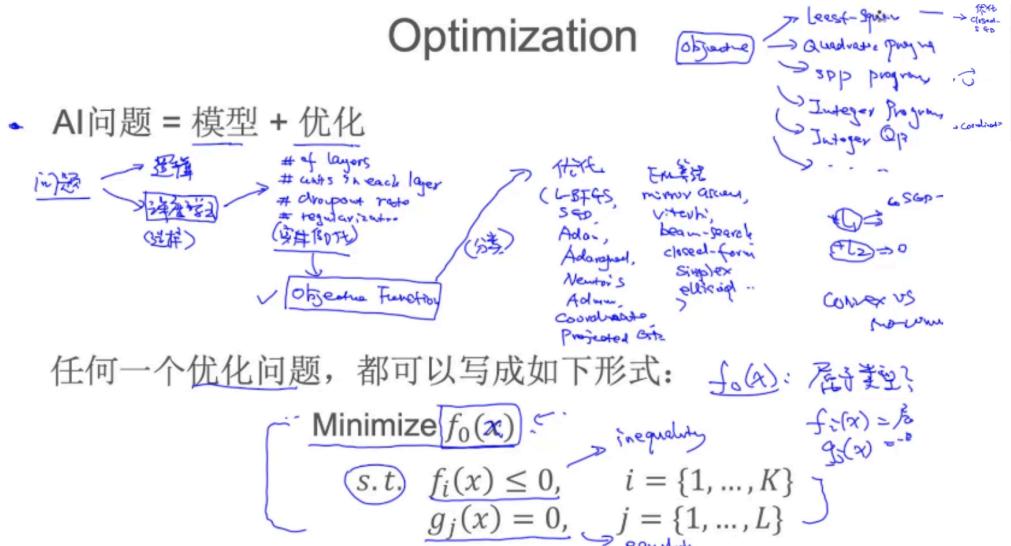


1. Optimization



任何一个优化问题，都可以写成如下形式： $f_0(x)$: 目标函数

$$\begin{aligned} & \text{Minimize } f_0(x) \\ & \text{s.t. } \begin{cases} f_i(x) \leq 0, & i = \{1, \dots, K\} \\ g_j(x) = 0, & j = \{1, \dots, L\} \end{cases} \end{aligned}$$

inequality

$f_0(x) = \beta$
 $f_i(x) = -\beta$
 $g_j(x) = 0$

Optimization is the Core of Machine Learning

线性回归(Linear Regression)

逻辑回归(Logistic Regression)

SVM(Support Vector Machine)

协同过滤(collaborative filtering)

K 均值(K-means)

1.1 categories

1) smooth / non-smooth

2) convex / non-convex

Convex 一定是全局最优解

Non-convex 是局部最优解, e.g. 深度学习

3) discrete / continuous

Discrete 一般需要用到离散数学

4) constrained / non-constrained

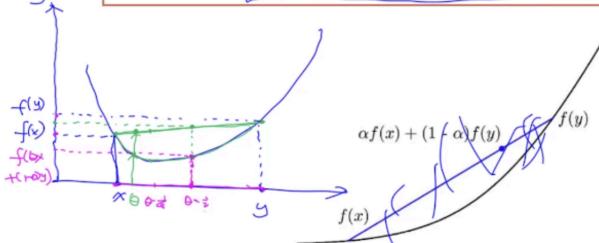
1.2 convex function

Convex Function (凸函数) $f(x)$ 是凸的
- $f(x)$ 是凸函数

凸函数定义

函数的定义域 $\text{dom } f$ 为凸集, 对于定义域里任意 x, y , 函数满足

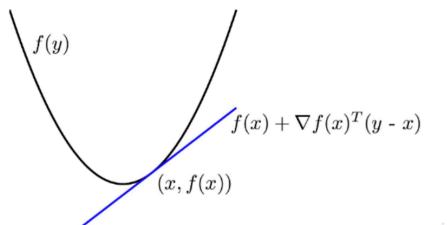
$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \quad \theta \in [0, 1]$$



- 线性函数为凸/凹函数
- $\exp x, -\log x, x \log x$ 是凸函数
- 范数为凸函数
- $\frac{x^T x}{t}$ 为凸函数 ($x > 0$)

First Order Convexity Condition

假设 $f: R^n \rightarrow R$ 是可导的 (differentiable), 则 f 为凸函数, 的当且仅当: $f(y) \geq f(x) + \nabla f(x)^T(y - x)$
对于任意 $x, y \in \text{dom}f$



Second Order Convexity Condition

假设 $f: R^n \rightarrow R$ 是两次可导的 (twice differentiable), 则 f 为凸函数, 当且仅当: $\nabla^2 f(x) \geq 0$

对于任意 $x, y \in \text{dom}f$

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

Scalar ≥ 0

Matrix \rightarrow PSD

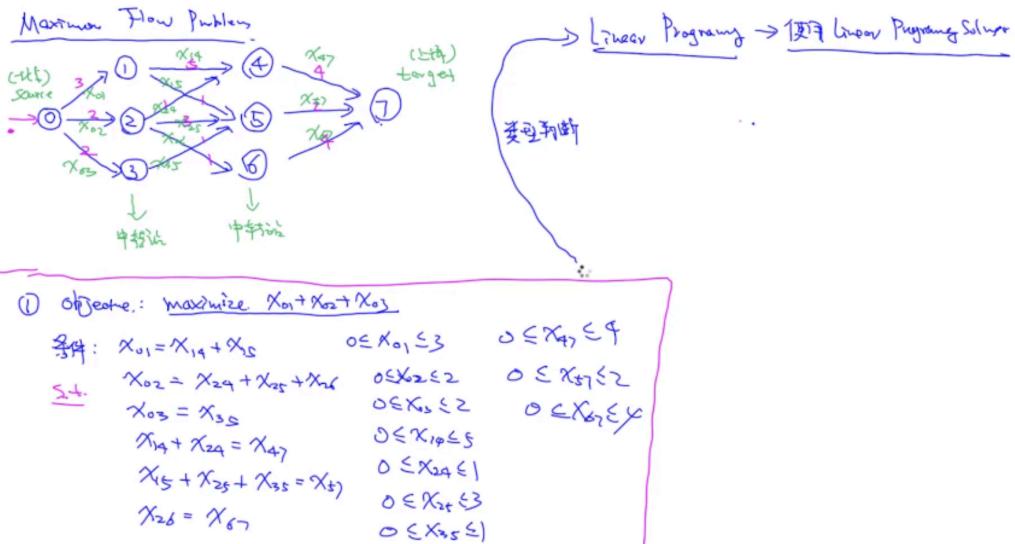
$\text{Positive} = \text{Semi-definite} \Rightarrow$
 $\text{正定} \Leftrightarrow$

- 线性函数: $f(x) = b^T x + c$

$$x_1, x_2 \Rightarrow f(x) = b^T \theta x_1 + c + f(x) = b^T x_2 + c.$$

$$\begin{aligned} \Rightarrow b^T (\theta x_1 + (1-\theta)x_2) + b^T c &\leq \theta(b^T x_1 + c) + (1-\theta)(b^T x_2 + c) \\ \theta b^T x_1 + (1-\theta)b^T x_2 + c &\leq \theta b^T x_1 + \theta c + (1-\theta)b^T x_2 + (1-\theta)c \end{aligned}$$

1.3 example-maximum flow problem



2. Set Cover Problem

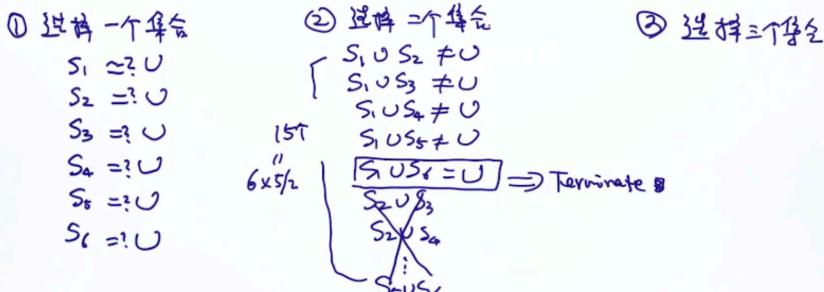
假设我们有个全集 U (Universal Set), 以及 m 个子集合 S_1, S_2, \dots, S_m , 目标是要 寻找最少的集合, 使得集合的 union 等于 U.

例子: $U = \{1, 2, 3, 4, 5\}$, $S: S_1 = \{1, 2, 3\}, S_2 = \{2, 4\}, S_3 = \{1, 3\}, S_4 = \{4\}, S_5 = \{3, 4\}, S_6 = \{4, 5\}$, 最少的集合为: {1,2,3}, {4,5}, 集合个数为2.

$$\{1, 2, 3\} \cup \{4, 5\} = \{1, 2, 3, 4, 5\} = U$$

2.1 Approach 1: Exhaustive Search

例子: $U = \{1, 2, 3, 4, 5\}$, $S: S_1 = \{1, 2, 3\}, S_2 = \{2, 4\}, S_3 = \{1, 3\}, S_4 = \{4\}, S_5 = \{3, 4\}, S_6 = \{4, 5\}$, 最少的集合为: {1,2,3}, {4,5}, 集合个数为2.

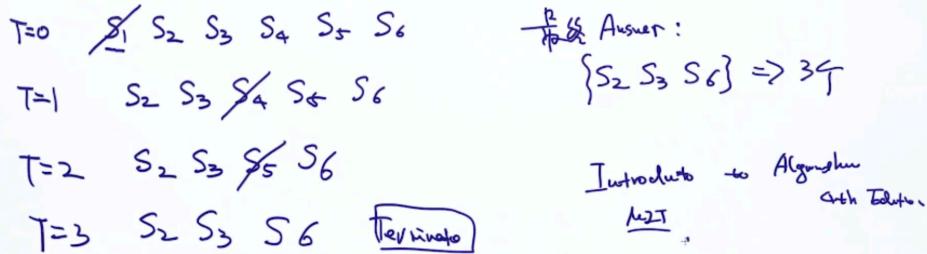


全局最优解

2.2 Approach 2: Greedy Search

Approach 2: Greedy Search \Rightarrow 不能保证全局最优解

例子: $U = \{1, 2, 3, 4, 5\}$, $S: \{S_1 = \{1, 2, 3\}, S_2 = \{2, 4\}, S_3 = \{1, 3\}, S_4 = \{4\}, S_5 = \{3, 4\}, S_6 = \{4, 5\}\}$, 最少的集合为: $\{1, 2, 3\}, \{4, 5\}$, 集合个数为2.



不能保证全局最优解

与穷举相比: 时间高效, 但是只是局部最优

2.3 Approach3: Optimization

Mathematical Formulation

$$\begin{aligned} m &= 6 \\ \text{例子: } U &= \{1, 2, 3, 4, 5\}, S: \{S_1 = \{1, 2, 3\}, S_2 = \{2, 4\}, S_3 = \{1, 3\}, S_4 = \{4\}, S_5 = \{3, 4\}, \\ &S_6 = \{4, 5\}\}, \text{最少的集合为: } \{1, 2, 3\}, \{4, 5\}, \text{集合个数为2.} \end{aligned}$$

目标: 设计什么样的算法?

$S_i \rightarrow x_i \in \{0, 1\}$ 当 $x_i = 1$ 时 \Rightarrow 选择 S_i $x_i = 0$ 时, 不选择 S_i

① 目标函数? $\sum_{i=1}^m x_i \leftarrow \text{minimize}$

② 条件? $x_i \in \{0, 1\} \in U, \sum_{i \in S_i} x_i \geq 1 \leftarrow \begin{cases} \text{if } i \in S_i \\ \text{else } 0 \end{cases} \Rightarrow$

\Rightarrow $\left\{ \begin{array}{l} \text{minimize } \sum_{i=1}^m x_i \\ \text{s.t. } \sum_{i \in S_i} x_i \geq 1, \forall i \in U \\ x_i \in \{0, 1\} \\ i = 1, 2, \dots, m \end{array} \right.$

Is it Convex?

$$\begin{aligned} &\text{minimize } \sum_{i=1}^m x_i \\ &\text{s.t. } \sum_{i \in S_i} x_i \geq 1 \\ &x_i \in \{0, 1\} \quad i = 1, \dots, m \end{aligned}$$

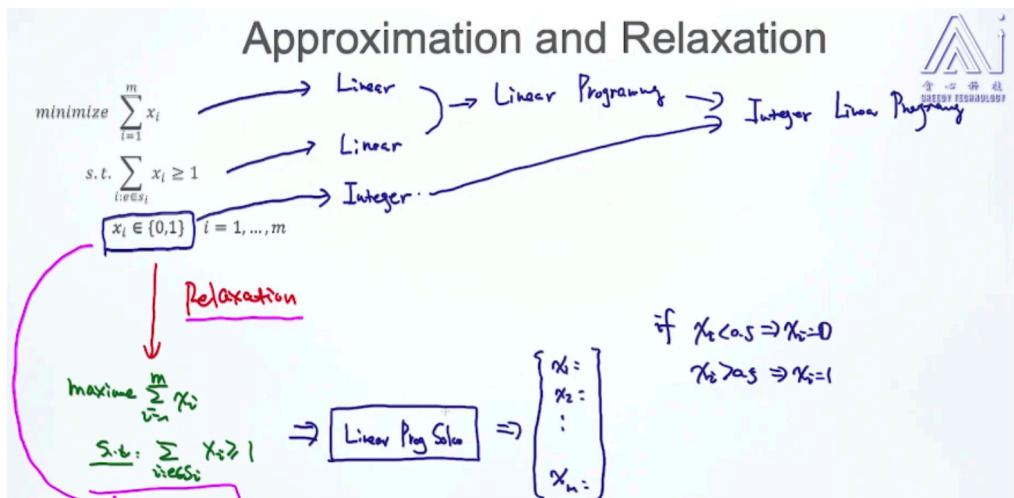
① 定义 $\frac{1}{2}$ Convex Set

$$\begin{matrix} \downarrow & \downarrow \\ 0 & 1 \end{matrix} \quad x_1, x_2 \quad \frac{\alpha x_1 + (1-\alpha)x_2}{2} \in \text{Convex Set}$$

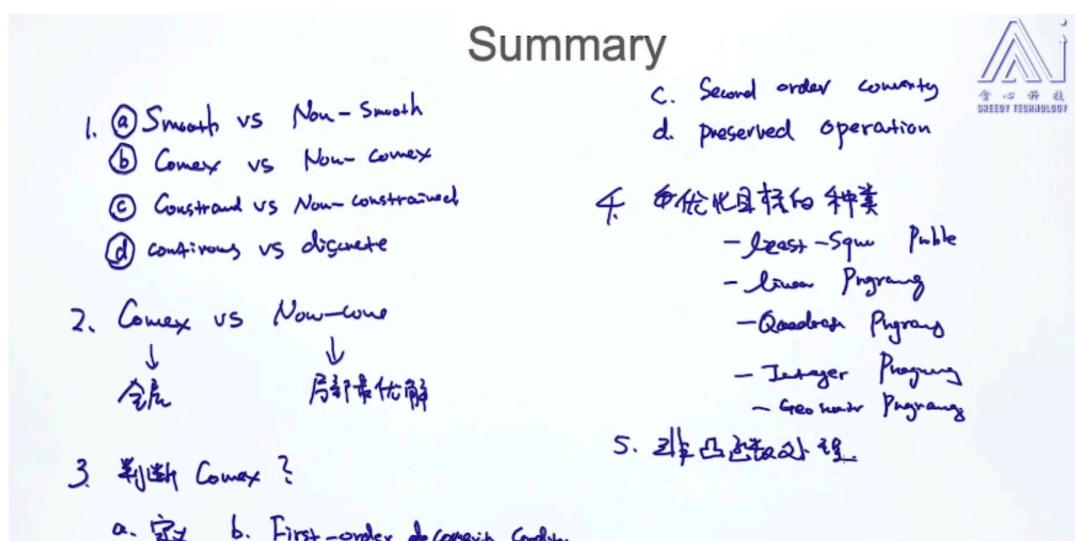


② 目标也需要 Convex

Not Convex!



3. summary



4. 梯度下降法

4.1 Recall: Gradient Descent for Logistic Regression

逻辑回归是一个凸函数

$$p(y=1|x, w) = \frac{1}{1 + e^{-w^T x + b}}$$

$$\operatorname{argmin}_{w,b} - \sum_{i=1}^n y \log p(y=1|x, w) + (1-y) \log(1-p(y=1|x, w))$$

$w^{t+1} = w^t - \eta_t \sum_{i=1}^n [\sigma(w^T x_i + b) - y_i] x_i$ $b^{t+1} = b^t - \eta_t \sum_{i=1}^n [\sigma(w^T x_i + b) - y_i]$
梯测 真实
Gradient Descent

4.2 梯度下降表达

梯度下降法的过程可以表示为: $f(x)$: x 是参数(变量)

1. 选择初始值 $x_0 \in R^d$ 和步长 (step-size) $\eta_t > 0$
2. $\text{for } i = 0, 1, \dots,$
 $x_{i+1} = x_i - \eta_t \nabla f(x_i)$

Iteration? Gd SGD

4.3 收敛分析 convergence analysis

定理 $f(x)$ 满足 L -Lipschitz 条件，并且是凸函数，设定 $x^* = \operatorname{argmin} f(x)$ ，那么对于步长 $\eta_t \leq \frac{1}{L}$ 满足常数。

x_k : 第 k 次迭代时的 x 值

$$f(x_k) \leq f(x^*) + \frac{\|x_0 - x^*\|_2^2}{2\eta_t k}$$

当我们迭代 $k = \frac{L\|x_0 - x^*\|_2^2}{\epsilon}$ 次之后我们可以保证得到 ϵ - approximation optimal value x ($\eta_t = 1/L$)

A $f(x_0) \leq f(x^*) + 1$
 $f(x_{10}) \leq f(x^*) + 0.01$
 B $f(x_0) \leq f(x^*) + 1$
 $f(x_{10}) \leq f(x^*) + 0.01$

当 k 变大 (迭代次数)，不等式右边第二项会变小。那么 $f(x_k)$ 会慢慢趋向于 $f(x^*)$ 。

4.4 凸函数性质回顾

定理: 任给 $x, y \in R^d$, $0 \leq \lambda \leq 1$

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

First-order Convexity:

$$f(x) + \nabla f(x)(y-x) \leq f(y) \quad ②$$

4.5 L-Lipschitz 条件及定理

定理一

一个光滑函数 (smooth function) f 满足 L -Lipschitz 条件，则对于任意 $x, y \in R^d$, 我们有：

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

Claim 1:

X: Tray Data
B730

i.e.: Linear regression.
 $L = \frac{1}{n} \sum_{i=1}^n \|Xw_i - y_i\|^2$
parameters

$$\begin{aligned} & \|\nabla f(w_1) - \nabla f(w_2)\| \\ &= \frac{2}{n} \left\| \nabla \cdot \left(\sum_{i=1}^n X^T (Xw_i - y_i) \right) \right\| \\ &= \frac{2}{n} \left\| \sum_{i=1}^n X^T X (w_1 - w_2) \right\| \leq \frac{2}{n} (\|X^T X\|) \cdot \|w_1 - w_2\| \end{aligned}$$

定理二

假设一个函数满足 L -Lipschitz 条件，并且是凸函数，对于任意 $x, y \in R^d$,

我们有: $f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{L}{2} \|y-x\|^2$ (Claim 2)

$$\begin{aligned} h(x) &= h(x) + \int_0^1 h'(z) dz \\ h(z) &= f(x+z(y-x)) \\ h(0) &= f(x), h(1) = f(y) \\ f(y) &= f(x) + \int_0^1 \nabla f(x+z(y-x))(y-x) dz \end{aligned}$$

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \nabla f(x+z(y-x))(y-x) dz \\ &= f(x) + \nabla f(x)(y-x) + \int_0^1 (\nabla f(x+z(y-x)) - \nabla f(x))(y-x) dz \\ &\leq f(x) + \nabla f(x)(y-x) + \int_0^1 L \|z(y-x)\| \|y-x\| dz \\ &= f(x) + \nabla f(x)(y-x) + \frac{L}{2} \|y-x\|^2 \end{aligned}$$

推导一

Derivation(1)

$$\begin{aligned} \text{Claim 2: } f(y) &\leq f(x) + \nabla f(x)(y-x) + \frac{L}{2} \|y-x\|^2 \\ f(x_{i+1}) &\leq f(x_i) + \nabla f(x_i)(x_{i+1} - x_i) + \frac{L}{2} \|x_{i+1} - x_i\|^2 \\ &= f(x_i) + \nabla f(x_i) - (-1) \cdot \eta_i \nabla f(x_i) + \frac{L}{2} \cdot \eta_i^2 \cdot \nabla f(x_i) \\ &= f(x_i) - \eta_i \|\nabla f(x_i)\|^2 + \frac{L \eta_i^2}{2} \|\nabla f(x_i)\|^2 \\ &= f(x_i) - \eta_i \left(1 - \frac{L \eta_i}{2}\right) \|\nabla f(x_i)\|^2 \\ &\leq f(x_i) - \frac{\eta_i}{2} \|\nabla f(x_i)\|^2 \\ f(x_{i+1}) &\leq f(x_i) - \frac{\eta_i}{2} \|\nabla f(x_i)\|^2 \end{aligned}$$

$x_{i+1} = x_i - \eta_i \nabla f(x_i)$

$\eta_i \leq \frac{1}{L}$

推导二

Derivation(2)

$$\begin{aligned}
 f(x_{i+1}) &\leq f(x_i) - \frac{\eta_t}{2} \|\nabla f(x_i)\|_2^2 \\
 &\leq f(x^*) + \nabla f(x_i)(x_i - x^*) - \frac{\eta_t}{2} \|\nabla f(x_i)\|_2^2 \quad \text{First Order Convexity} \\
 &= f(x^*) + \frac{x_i - x_{i+1}}{\eta_t} \cdot (x_i - x^*) - \frac{1}{2\eta_t} \|x_i - x_{i+1}\|^2 \\
 &= f(x^*) + \frac{1}{2\eta_t} \|x_i - x^*\|^2 - \frac{1}{2\eta_t} (\|x_i - x^*\|^2 - 2\eta_t \nabla f(x_i)(x_i - x^*) + \|\nabla f(x_i)\|^2) \\
 &= f(x^*) + \frac{1}{2\eta_t} \|x_i - x^*\|^2 - \frac{1}{2\eta_t} \|x_i - x^* - \nabla f(x_i)\|^2 \quad x_i - x_{i+1} \\
 &= f(x^*) + \frac{1}{2\eta_t} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2)
 \end{aligned}$$

$$f(x_{i+1}) \leq f(x^*) + \frac{1}{2\eta_t} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2)$$

推导三

Derivation(3)

$$\begin{aligned}
 f(x_{i+1}) &\leq f(x^*) + \frac{1}{2\eta_t} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2) \\
 f(x_{i+1}) - f(x^*) &\leq \frac{1}{2\eta_t} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2) \\
 \oplus \left\{ \begin{array}{l} f(x_1) - f(x^*) \leq \frac{1}{2\eta_t} (\|x_0 - x^*\|^2 - \|x_1 - x^*\|^2) \\ f(x_2) - f(x^*) \leq \frac{1}{2\eta_t} (\|x_1 - x^*\|^2 - \|x_2 - x^*\|^2) \\ f(x_3) - f(x^*) \leq \frac{1}{2\eta_t} (\|x_2 - x^*\|^2 - \|x_3 - x^*\|^2) \\ \vdots \quad \vdots \\ f(x_k) - f(x^*) \leq \frac{1}{2\eta_t} (\|x_{k-1} - x^*\|^2 - \|x_k - x^*\|^2) \end{array} \right. \\
 \sum_{i=1}^k f(x_i) - k \cdot f(x^*) &\leq \frac{1}{2\eta_t} (\|x_0 - x^*\|^2 - \|x_k - x^*\|^2) \\
 \sum_{i=1}^k f(x_i) - k \cdot f(x^*) &\leq \frac{1}{2\eta_t} \|x_0 - x^*\|^2
 \end{aligned}$$

$$\begin{aligned}
 f(x_{i+1}) &\leq f(x_i) - \frac{\eta_t}{2} \|\nabla f(x_i)\|_2^2 \\
 f(x_{i+1}) &\leq f(x_i) \leq f(x_{i+1}) \dots \leq f(x^*) \\
 k \cdot f(x_i) - k \cdot f(x^*) &\leq \frac{k}{2\eta_t} f(x_i) + f(x^*) \\
 &\leq \frac{1}{2\eta_t} \|x_0 - x^*\|^2 \\
 \downarrow \\
 k \cdot f(x_i) - k \cdot f(x^*) &\leq \frac{1}{2\eta_t} (\|x_0 - x^*\|^2) \\
 \boxed{f(x_i) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\eta_t \cdot k}}
 \end{aligned}$$