Minicurso de verão da FGV/2025
Fundamentos da Otimização Multiobjetivo: **Ponto Proximal**

Jefferson D.G. Melo

IME/UFG

15 de janeiro de 2025

Descrição desta aula

- O problema de otimização multiobjetiva
- Revisão do método do ponto proximal para função escalar
- Método do ponto proximal multiobjetivo
- Análise de convergência do método ponto proximal multiobjetivo
- Discussão sobre o método proximal-gradiente multiobjetivo

O problema de otimização multiobjetivo e solução Pareto ótima

- Nesta parte estamos interessados no seguinte problema convexo multiobjetivo:

$$\min_{x\in\mathbb{R}^n} F(x), \tag{1}$$

  sendo $F : \mathbb{R}^n \to \mathbb{R}^m$ uma função multiobjetiva convexa, não necessariamente diferenciável.

- F é convexa se $F(\alpha x + (1-\alpha)y) \preceq \alpha F(x) + (1-\alpha)F(y)$ para todo $x, y \in \mathbb{R}^n$ e $\alpha \in [0,1]$.
- Um ponto $x^* \in \mathbb{R}^n$ é uma solução Pareto fraca de (1) se **não existe** $x \in \mathbb{R}^n$ tal que $F(x) \prec F(x^*)$.
- Utilizamos a ordem parcial usual em $\mathbb{R}^m$: se $a, b \in \mathbb{R}^m$, temos

$$a \preceq b \Longleftrightarrow a_i \leq b_i \quad \text{e} \quad a \prec b \Longleftrightarrow a_i < b_i, \quad \forall i \in \{1,\dots,m\}. \tag{2}$$

Ilustração de uma solução Pareto e Pareto fraco

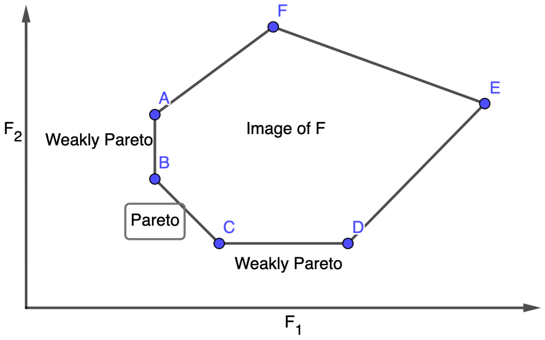
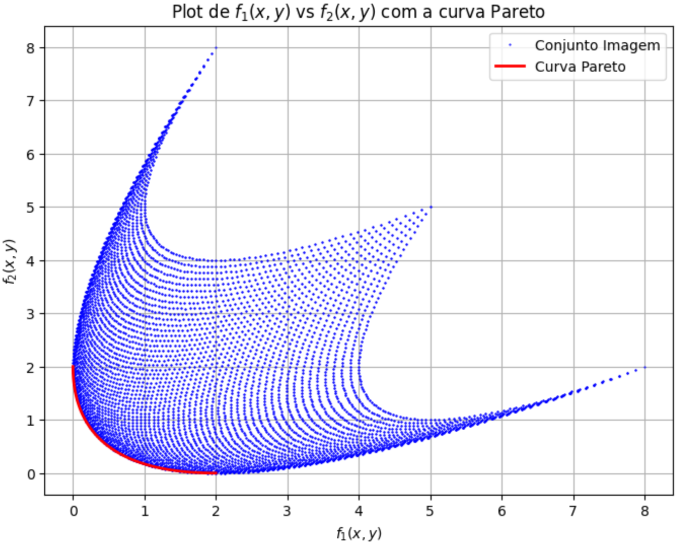


Ilustração de curva Pareto



Condição de otimalidade para problema escalar e convexo

Considere o seguinte problema de otimização escalar:

$$\min_{x \in \mathbb{R}^n} \; f(x),$$

onde $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa não necessariamente diferenciável.

Uma condição necessária e suficiente para que um ponto $x^* \in \mathbb{R}^n$ seja uma solução ótima do problema acima é:

$$0 \in \partial f(x^*),$$

onde $\partial f(x)$ é o subdiferencial de f em x, definido por

$$\partial f(x) = \{v \in \mathbb{R}^n : \; f(y) \geq f(x) + \langle v, y - x \rangle, \quad \forall y \in \mathbb{R}^n\}.$$

Subdiferencial de funções convexas

Seja $f : \mathbb{R}^n \to \mathbb{R}$ convexa, então o subdiferencial de f em x é definido por

$$\partial f(x) := \{v \in \mathbb{R}^n : f(y) \geq f(x) + \langle v, y - x \rangle \,, \ \forall y \in \mathbb{R}^n\} \tag{3}$$

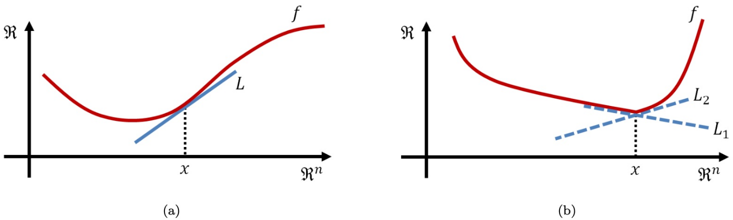
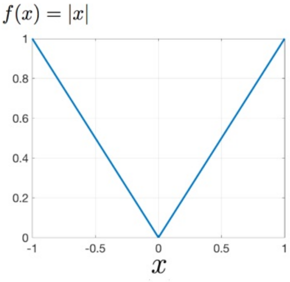


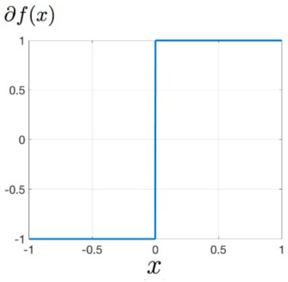
Figura: a) Não convexa, b) Função convexa e subgradientes em x

Note que $\bar{x}$ é um minimizador de f se e somente se $0 \in \partial f(\bar{x})$.

Subdiferencial de $f(x) := |x|$

$$\partial f(x) := \{v \in \mathbb{R}^n : f(y) \geq f(x) + \langle v, y - x \rangle, \ \forall y \in \mathbb{R}^n\}$$





$$f(x) = |x| \qquad \partial f(x) = \begin{cases} \{-1\}, & \text{if } x < 0 \\ [-1, 1], & \text{if } x = 0 \\ \{1\}, & \text{if } x > 0 \end{cases}$$

Crescimento quadrático de uma função fortemente convexa

$f : \mathbb{R}^n \to \mathbb{R}$ é fortemente convexa com parâmetro μ se

$$f(x + \alpha(y - x)) \leq f(x) + \alpha(f(y) - f(x)) - \frac{\mu}{2}\alpha(1 - \alpha)\|x - y\|^2, \quad \forall \alpha \in [0, 1].$$

Assuma que x seja minimizador de f sobre um conjunto convexo C. Então, para todo $y \in C$:

$$0 \leq f(x + \alpha(y - x)) - f(x) \leq \alpha(f(y) - f(x)) - \frac{\mu}{2}\alpha(1 - \alpha)\|x - y\|^2, \quad \forall \alpha \in [0, 1].$$

Dividindo por $\alpha > 0$ e passando o limite com $\alpha \to 0$, obtemos

$$0 \leq f(y) - f(x) - \frac{\mu}{2}\|x - y\|^2,$$

ou seja

$$f(y) \geq f(x) + \frac{\mu}{2}\|y - x\|^2, \quad \forall y \in C.$$

Método do ponto proximal

O método do ponto proximal é uma abordagem clássica e fundamental na resolução de problemas convexos não necessariamente suaves.
Características importantes incluem:

- A robustez em resolver problemas de otimização cuja função objetiva seja convexa possivelmente não diferenciável;
- Seus subproblemas podem possuir soluções explicitas em algumas aplicações importantes, como o problema do lasso, por exemplo;
- Conexão com vários outros métodos modernos em otimização, incluindo métodos de regularização, Lagrangiano aumentado, etc...

Método do ponto proximal (escalar)

Formalmente, o método consiste em, a partir de um ponto inicial x^0, gerar uma sequência $\{x^k\}$ de pontos por meio do seguinte procedimento iterativo:

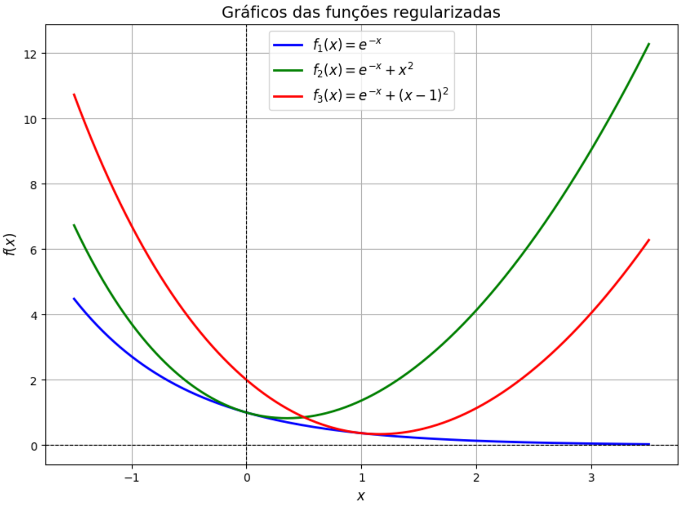
$$x^{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\lambda_k} \|x - x^k\|^2 \right\}, \tag{4}$$

onde $\lambda_k > 0$ é um parâmetro de regularização que pode variar a cada iteração.
Essa formulação transforma o problema original em uma sequência de problemas regularizados. De fato, note que a função objetiva do subproblema proximal é a soma de uma função convexa com uma fortemente convexa, logo ela também é fortemente convexa e possui um único minimizador x^{k+1}. Da condição de otimalidade de x^{k+1}, temos a maneira equivalente de formular o método por meio do problema de inclusão monótona:

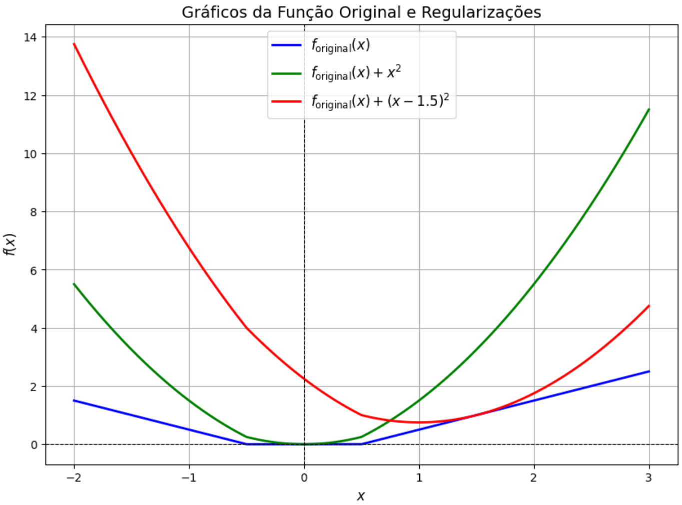
$$0 \in \partial f(x^{k+1}) + \frac{1}{\lambda_k}(x^{k+1} - x^k).$$

Em particular, podemos usar como critério de parada a verificação se $x^{k+1} = x^k$, isto implicaria $0 \in \partial f(x^{k+1})$, e portanto que x^{k+1} seja uma solução.

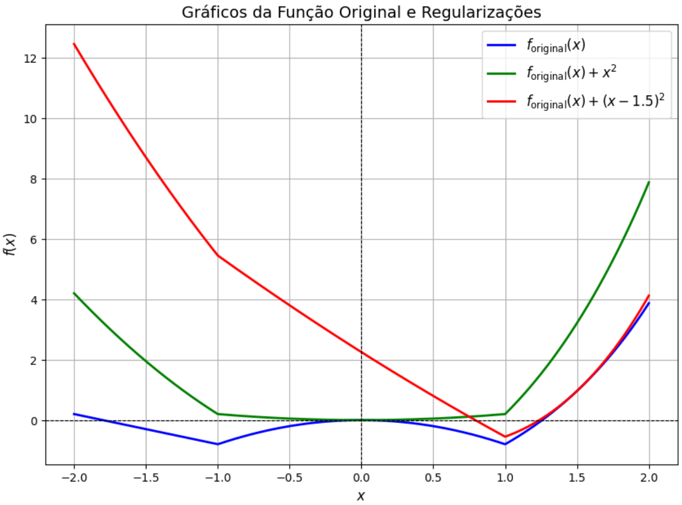
Função sem minimizador, regularizada passa a ter um minimizador



Função com vários mínimos, regularizada tem um único minimizador



Função regularizada pode convexificar a função original



Método do Ponto Proximal para Otimização Multiobjetivo

Agora retornamos ao problema de otimização multiobjetivo:

$$\min_{x \in \mathbb{R}^n} \; F(x) = (F_1(x), F_2(x), \ldots, F_m(x)),$$

onde $F : \mathbb{R}^n \to \mathbb{R}^m$ é uma função convexa (não necessariamente diferenciável).
A ideia principal no contexto multiobjetivo é buscar por pontos de Pareto fraco, ou seja, soluções onde não é possível melhorar todas as funções componentes. O método do ponto proximal multiobjetivo adapta-se para encontrar tais pontos e suas características importantes incluem:

- O método é um esquema geral que independe de escalarizações, e ainda assim mantêm a conhecida robustez do método proximal escalar;
- Os subproblemas regularizam todos as componentes da função multiobjetiva, garantindo descida em todas as componentes e a convergência para soluções Pareto fracas.

Formulação do método do ponto proximal multiobjetivo

O método do ponto proximal multiobjetivo[1], pode ser visto como um esquema geral que produz uma sequência de pontos $\{x^k\}$ por meio do seguinte procedimento iterativo:

Compute x^{k+1} como sendo uma solução Pareto fraca do seguinte problema multiobjetivo

$$\min_{x \in \Omega_k} \left\{ F(x) + \frac{1}{2\lambda_k} \|x - x^k\|^2 e^k \right\}, \tag{5}$$

onde $\Omega_k := \{x : F(x) \preceq F(x^k)\}$, $\lambda_k > 0$ é um parâmetro de regularização, e^k é vetor unitário de coordenadas positivas tal $\inf_{\{z \succeq 0 : \|z\|=1\}} \langle e^k, z \rangle \geq \delta$ para todo $k \geq 0$ e algum $\delta > 0$.
Note que o método não especifica como computar sua solução Pareto fraca do subproblema. Para isso, pode ser utilizado diferentes algoritmos ou estratégias, como por exemplo, escalarizações lineares ou do máximo.

[1] Bonnel, H., Iusem, A.N., Svaiter, B.F.: Proximal methods in vector optimization. SIAM J. Optim. 15(4):953-970, 2005.

Solução Pareto fraca do subproblema proximal

Note que vale a seguinte relação[2]

$$\arg\min_{x\in C} G(x) = \bigcup_{z\succeq 0} \{\arg\min_{x\in C} \langle G(x), z\rangle\}.$$

Logo, para cada $k \geq 0$, temos que existe $z^k \succeq 0$, não nulo, tal que

$$x^{k+1} = \arg\min_{x\in\Omega_k} \left\{ \langle F(x), z^k\rangle + \frac{1}{2\lambda_k}\langle e^k, z^k\rangle \|x - x^k\|^2 \right\}. \tag{6}$$

Como o conjunto de nível $\Omega_k = \{x \in \mathbb{R}^n : F(x) \preceq F(x^k)\}$ é convexo e a função objetiva do subproblema acima é fortemente convexa, temos que existe um único ponto x^{k+1} associado ao vetor z^k.

[2] [Theorems 2.10-2.11] T. D. Luc, Theory of Vector Optimization, Lecture Notes in Econom. and Math. Syst. 319, Springer, Berlin, 1989.

Condição de otimalidade do subproblema proximal

Temos pela condição de otimalidade[3] de x^{k+1} e convexidade de F que

$$0 \in \partial \langle z^k, F \rangle (x^{k+1}) + \frac{1}{\lambda_k} \langle e^k, z^k \rangle (x^{k+1} - x^k) + \mathcal{N}_{\Omega_k}(x^k).$$

- Podemos assumir, sem perda de generalidades que $\|z^k\| = 1$. Além disso, existe $\delta > 0$ tal que $\langle e^k, z^k \rangle \geq \delta$ para todo $k \geq 0$.
- Por conveniência, denotaremos por $\eta_k := \frac{1}{\lambda_k} \langle e^k, z^k \rangle$, e assumiremos que $\{\lambda_k\}$ é limitada. Podemos então assumir sem perda de generalidade que η_k converge para um certo $\bar{\eta}$.

[3] $\mathcal{N}_C(x)$ é o cone normal a C no ponto x.

Outras propriedades importantes

Se x^k é um Pareto fraco de F, como $x^k \in \Omega_k$, temos que existe $z^k \succeq 0$, não nulo, tal que

$$x^k = \arg \min_{x \in \Omega_k} \langle F(x), z^k \rangle,$$

o qual implica que para todo $x \in \Omega_k$, temos

$$\langle F(x^k), z^k \rangle + \frac{1}{2\lambda_k} \langle e^k, z^k \rangle \|x^k - x^k\|^2 = \langle F(x^k), z^k \rangle \leq \langle F(x), z^k \rangle \tag{7}$$

$$\leq \langle F(x), z^k \rangle + \frac{1}{2\lambda_k} \langle e^k, z^k \rangle \|x - x^k\|^2, \tag{8}$$

concluindo portanto que

$$\text{se } x^k \in \arg \min F(x) \Rightarrow x^k \in \arg \min_{x \in \Omega_k} F(x) + \frac{1}{2\lambda_k} \|x - x^k\| e^k,$$

i.e., x^k é uma sulução Pareto fraca do subproblema proximal.

Critério de parada alternativo: $x^{k+1} = x^k$

Note que a relação acima não implica que $x^{k+1} = x^k$, ela implica que se computamos x^{k+1} utilizando uma escalarização z^k "especial", então aí sim temos que $x^k = x^{k+1}$.
Assuma agora que $x^{k+1} = x^k$. Então temos da condição de otimalidade do subproblema proximal que existe $z^k \succeq 0$, não nulo, tal que

$$0 \in \partial\langle z^k, F\rangle(x^{k+1}) + \eta_k(x^{k+1} - x^k) + \mathcal{N}_{\Omega_k}(x^{k+1})$$

e portanto, como $x^{k+1} = x^k$, temos

$$0 \in \partial\langle z^k, F\rangle(x^k) + \mathcal{N}_{\Omega_k}(x^k)$$

o que implica que x^k minimiza $\langle z^k, F(x)\rangle$ em Ω_k, e portanto é Pareto fraco de $F(x)$ em Ω_k. Mas note que isso implica que não existe $x \in \Omega_k$ tal que $F(x) \prec F(x^k)$. Em particular, temos que não existe nenhum $x \in \mathbb{R}^n$ tal que $F(x) \prec F(x^k)$, pois caso contrário, esse ponto x estaria em Ω_k e contrariaria a conclusão que x^k é Pareto fraco em Ω_k.

A sequência $\{\|x^{k+1} - x^k\|^2\}$ é somável

- Como $x^{k+1} \in \Omega_k$, temos que $F(x^{k+1}) \preceq F(x^k)$, logo $\{F(x^k)\}$ é decrescente (na ordem parcial de $\mathbb{R}^m$), e portanto ela converge monotonicamente para um certo $\tilde{F}$;
- Do subproblema proximal, observando que $x^k \in \Omega_k$, temos:

$$\langle F(x^{k+1}), z^k\rangle + \eta_k \|x^{k+1} - x^k\|^2 \leq \langle F(x^k), z^k\rangle.$$

De onde segue que

$$\eta_k \|x^{k+1} - x^k\|^2 \leq \langle F(x^k) - F(x^{k+1}), z^k\rangle.$$

Agora note que $\langle e^k, z^k\rangle \geq \delta$, $\lambda_k \leq \hat{\lambda}$ e $F_j(x^k) - F_j(x^{k+1}) \geq 0$ e $0 \leq z_j^k \leq \|z^k\| = 1$. Em particular, temos

$$\eta_k = \frac{\langle e^k, z^k\rangle}{2\lambda_k} \geq \frac{\delta}{2\hat{\lambda}}.$$

Continuação da prova

Portanto, temos da desigualdade anterior que

$$
\begin{aligned}
\frac{\delta}{2\hat{\lambda}} \sum_{k=0}^{N} \|x^{k+1} - x^k\|^2 &\leq \sum_{k=0}^{N} \langle F(x^k) - F(x^{k+1}), z^k \rangle = \sum_{k=0}^{N} \sum_{j=1}^{m} \left(F_j(x^k) - F_j(x^{k+1}) \right) z_j^k \\
&\leq \sum_{k=0}^{N} \sum_{j=1}^{m} \left(F_j(x^k) - F_j(x^{k+1}) \right) \\
&= \sum_{j=1}^{m} \left(F_j(x^0) - F_j(x^{N+1}) \right) \leq \sum_{j=1}^{m} \left(F_j(x^0) - \tilde{F}_j \right).
\end{aligned}
$$

Concluímos dessas desigualdades que $\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\|^2 < +\infty$. Em particular, a sequência $\{\|x^{k+1} - x^k\|\}$ converge para zero.

Definição e resultado básico de Fejér convergência

Seja S um subconjunto não vazio de $\mathbb{R}^n$. Uma sequência $\{x^k\}$ em $\mathbb{R}^n$ é dita Fejér convergente para S se, e somente se, para todo $x \in S$, existe uma sequência somável $\{\varepsilon_k\} \subset \mathbb{R}_+$ tal que, para todo inteiro $k \geq 0$, temos

$$\|x^{k+1} - x\|^2 \leq \|x^k - x\|^2.$$

A importância da convergência Fejér é ilustrada no seguinte lema.

Lema: Se $\{x^k\}$ é Fejér convergente para um conjunto não vazio S, então as seguintes afirmações são válidas:

(i) a sequência $\{x^k\}$ é limitada;

(ii) se $\{x^k\}$ tem ponto de acumulação em S, então $\{x^k\}$ converge para um ponto em S.

Fejér convergência do método ponto proximal multiobjetivo

Teorema: Assuma que $\Omega := \cap_k \Omega_k \neq \emptyset$ e que $0 < \lambda_k \leq \hat{\lambda}$. Então, a sequência $\{x^k\}$ gerada pelo método proximal multiobjetivo converge a um ponto Pareto fraco de F.

Fejér convergência:
Utilizando a otimalidade de x^{k+1} e o fato que a função objetiva do subproblema proximal é fortemente convexa, temos para todo $x \in \Omega \subset \Omega_k$:

$$\langle F(x), z^k \rangle + \eta_k \|x - x^k\|^2 \geq \langle F(x^{k+1}), z^k \rangle + \eta_k \|x^{k+1} - x^k\|^2 + \eta_k \|x - x^{k+1}\|^2. \tag{9}$$

Como $\langle F(x^{k+1}), z^k \rangle \geq \langle F(x), z^k \rangle$ e $\eta_k > 0$, a desigualdade acima implica que

$$\|x - x^k\|^2 \geq \|x - x^{k+1}\|^2 + \|x^{k+1} - x^k\|^2. \tag{10}$$

Concluindo que $\{x^k\}$ é Fejér convergente a Ω, como todo ponto de acumulação de $\{x^k\}$ pertence a Ω (pois $\{F(x^k)\}$ decresce), concluímos que a sequência $\{x^k\}$ converge a um ponto de Ω. obs.: Note que (10) também implica que $\|x^{k+1} - x^k\|^2$ é somável (neste caso, utilizando que $\Omega \neq 0$.

Continuação da análise de convergência

Vamos mostrar agora que o ponto limite $\bar{x}$ de $\{x^k\}$ é uma solução Pareto fraca de F. Assuma por contradição que, $\bar{x}$ não seja Pareto fraco de F. Logo, existe $x \in \mathbb{R}^n$ tal que $F(x) \prec F(\bar{x})$. Como $\{F(x^k)\}$ decresce, temos que $F(\bar{x}) \preceq F(x^k)$, para todo $k \geq 0$, ou seja, $\bar{x} \in \Omega_k$ e consequentemente, x também pertence a Ω_k, pois $F(x) \prec F(\bar{x})$. Logo da otimalidade de x^{k+1} e do fato que a função do subproblem proximal é fortemente convexa, temos

$$\langle F(x^{k+1}), z^k \rangle + \eta_k \|x^{k+1} - x^k\|^2 + \eta_k \|x - x^{k+1}\|^2 \leq \langle F(x), z^k \rangle + \eta_k \|x - x^k\|^2. \qquad (11)$$

Note que $\{\eta_k\}$ converge a $\bar{\eta}$ e $\{\|x^{k+1} - x^k\|\}$ converge para zero. Além disso, $\{z^k\}$ é limitada e portanto podemos assumir sem perda de generalidade que ela converge para um certo $\bar{z} \succeq 0$, não nulo. Logo, passando limite com $k \to +\infty$ em ambos os lados da desigualdade acima:

$$\langle F(\bar{x}), \bar{z} \rangle + \bar{\eta} \|x - \bar{x}\|^2 \leq \langle F(x), \bar{z} \rangle + \bar{\eta} \|x - \bar{x}\|^2,$$

ou seja$\langle F(\bar{x}), \bar{z} \rangle \leq \langle F(x), \bar{z} \rangle$, o que contradiz o fato que $F(x) \prec F(\bar{x})$.