

# Convex Optimization with Inexact Model of the Objective Function

Alexander Tyurin <sup>1</sup>, Pavel Dvurechensky <sup>2</sup>,  
**Alexander Gasnikov** <sup>3</sup>

<sup>1</sup>HSE (Russian Federation)

<sup>2</sup>WIAS (Germany)

<sup>3</sup>MIPT (Russian Federation)

# General optimization problem

General convex optimization problem:

$$F(x) \rightarrow \min_{x \in Q}$$

We denote  $x^*$  – the solution of this problem.

- 1  $F(x)$  is a convex function.
- 2  $Q \subseteq \mathbb{R}^n$  is a convex, closed set.

**Examples:**

- 1 Physics, Gas Network:

$$\frac{1}{3} \sum_{i=1}^N \alpha_i |x_i|^3 \rightarrow \min_{Ax=d} .$$

- 2 Machine Learning, Logistic regression:

$$\sum_{i=1}^N \log(1 + \exp(x_i^T w)) \rightarrow \min_{x \in \mathbb{R}^n} .$$

**Prox-function**  $d(x)$  : continuously differentiable on  $\text{int}Q$  and 1-strongly convex function, i.e.

$$d(x) \geq d(y) + \langle \nabla d(y), x - y \rangle + \frac{1}{2} \|x - y\|^2, \quad x, y \in \text{int} Q.$$

**Bregman divergence**  $V(x, y) \stackrel{\text{def}}{=} d(x) - d(y) - \langle \nabla d(y), x - y \rangle$ .

**Examples:**

- $V(x, y) = \frac{1}{2} \|x - y\|_2^2$
- $V(x, y) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i}$  (KL-divergence).

# Definition of $(\delta, L)$ -model

## Definition

Pair  $(F_\delta(y), \psi_\delta(x, y))$  is  $(\delta, L)$ -model in a point  $y$  of a function  $F$  with respect to the norm  $\|\cdot\|$  if for all  $x \in Q$

$$0 \leq F(x) - F_\delta(y) - \psi_\delta(x, y) \leq \frac{L}{2} \|x - y\|^2 + \delta,$$

$$\psi_\delta(x, x) = 0, \quad \forall x \in Q,$$

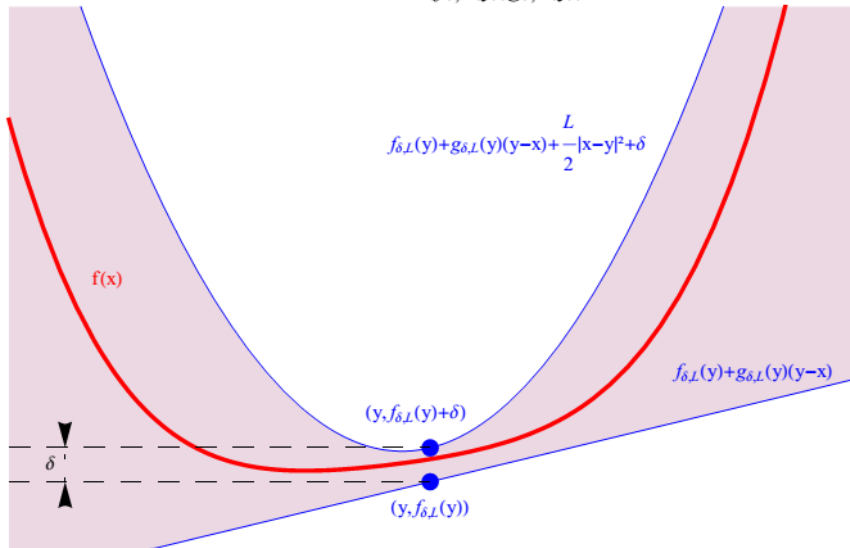
and  $\psi_\delta(x, y)$  is a convex function of  $x$  for all  $y \in Q$ .

## Definition (Devolder–Glineur–Nesterov, 2013)

Function  $F$  has  $(\delta, L)$ -oracle in a point  $y$  if there exists a pair  $(F_\delta(y), \nabla F_\delta(y))$  such that

$$0 \leq F(x) - F_\delta(y) - \langle \nabla F_\delta(y), x - y \rangle \leq \frac{L}{2} \|x - y\|^2 + \delta, \quad \forall x \in Q.$$

Inexact oracle  $(f_{\delta,L}(y), g_{\delta,L}(y))$



## Definition

Let us assume an optimization task

$$\psi(x) \rightarrow \min_{x \in Q},$$

where  $\psi$  is a convex function. Then  $\text{Arg min}_{x \in Q}^{\tilde{\delta}} \psi(x)$  is a set of  $\tilde{x}$  such that

$$\exists h \in \partial\psi(\tilde{x}), \quad \langle h, x_* - \tilde{x} \rangle \geq -\tilde{\delta}.$$

Arbitrary point from the  $\text{Arg min}_{x \in Q}^{\tilde{\delta}} \psi(x)$  we define as  $\arg \min_{x \in Q}^{\tilde{\delta}} \psi(x)$ .

From the definition follows that

$$\psi(\tilde{x}) - \psi(x_*) \leq \tilde{\delta}.$$

In the opposite direction, it is not always true.

# Gradient descent

**Input:**  $x_0$ ,  $N$  – number of steps,  $\{\delta_k\}_{k=0}^{N-1}$ ,  $\{\tilde{\delta}_k\}_{k=0}^{N-1}$  – sequences, and  $L_0 > 0$ .

$$\mathbf{0 - step: } L_1 = \frac{L_0}{2}$$

$$\mathbf{k + 1 - step: } \alpha_{k+1} = \frac{1}{L_{k+1}}$$

$$\phi_{k+1}(x) = V(x, x_k) + \alpha_{k+1} \psi_{\delta_k}(x, x_k)$$

$$x_{k+1} = \arg \min_{x \in Q}^{\tilde{\delta}_k} \phi_{k+1}(x)$$

$$\mathbf{If } F_{\tilde{\delta}_k}(x_{k+1}) > F_{\tilde{\delta}_k}(x_k) + \psi_{\delta_k}(x_{k+1}, x_k) + \frac{L_{k+1}}{2} \|x_{k+1} - x_k\|^2 + \delta_k, \mathbf{ then } L_{k+1} = 2L_{k+1}$$

**and repeat the current step.**

**Otherwise,**  $L_{k+2} = \frac{L_{k+1}}{2}$  **and go to the next step.**

# Fast gradient descent

**Input:**  $x_0$ ,  $N$  – number of steps,  $\{\delta_k\}_{k=0}^{N-1}$ ,  $\{\tilde{\delta}_k\}_{k=0}^{N-1}$  – sequences, and  $L_0 > 0$ .

**0 - step:**  $y_0 = u_0 = x_0$ ;  $L_1 = \frac{L_0}{2}$ ;  $\alpha_0 = 0$ ;  $A_0 = \alpha_0$

**$k + 1$  - step: Find  $\alpha_{k+1}$  :**  $A_k + \alpha_{k+1} = L_{k+1}\alpha_{k+1}^2$

$$A_{k+1} = A_k + \alpha_{k+1}$$

$$y_{k+1} = \frac{\alpha_{k+1}u_k + A_k x_k}{A_{k+1}}$$

$$\phi_{k+1}(x) = V(x, u_k) + \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1})$$

$$u_{k+1} = \arg \min_{x \in Q} \tilde{\delta}_k \phi_{k+1}(x), \quad x_{k+1} = \frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}$$

**If  $F_{\delta_k}(x_{k+1}) > F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_{k+1}, y_{k+1}) + \frac{L_{k+1}}{2} \|x_{k+1} - y_{k+1}\|^2 + \delta_k$ , then  $L_{k+1} = 2L_{k+1}$**

**and repeat the current step.**

**Otherwise,  $L_{k+2} = \frac{L_{k+1}}{2}$  and go to the next step.**



## Theorem (Gradient descent)

Let  $V(x_*, x_0) \leq R^2$  where  $x_0$  is a starting point,  $x_*$  is the closest point to  $x_0$  in terms of a Bregman distance  $V$ , and

$$\bar{x}_N = \frac{1}{N} \sum_{k=0}^{N-1} x_k.$$

For proposed algorithm we have

$$F(\bar{x}_N) - F(x_*) \leq \frac{2LR^2}{N} + \frac{2L}{N} \sum_{k=0}^{N-1} \tilde{\delta}_k + \frac{2}{N} \sum_{k=0}^{N-1} \delta_k.$$

## Theorem (Fast gradient descent)

Let  $V(x_*, x_0) \leq R^2$  where  $x_0$  is a starting point and  $x_*$  is the closest point to  $x_0$  in terms of a Bregman distance  $V$ . For proposed algorithm we have

$$F(x_N) - F(x_*) \leq \frac{8LR^2}{(N+1)^2} + \frac{8L \sum_{k=0}^{N-1} \tilde{\delta}_k}{(N+1)^2} + \frac{2 \sum_{k=0}^{N-1} \delta_k A_{k+1}}{A_N}.$$

## Gradient descent vs Fast gradient descent

Let  $\tilde{\delta}_k = \tilde{\delta}$  and  $\delta_k = \delta$  then

$$F(\bar{x}_N) - F(x_*) \leq \frac{2LR^2}{N} + 2L\tilde{\delta} + 2\delta,$$

$$F(x_N) - F(x_*) \leq \frac{8LR^2}{(N+1)^2} + \frac{8L\tilde{\delta}}{N+1} + 2N\delta.$$

Fast gradient descent is more robust to  $\tilde{\delta}$ , which appears when we solve an intermediate optimization problem, but fast gradient descent accumulates  $\delta$ , which appears when we receive  $(\delta, L)$ -model.

There are intermediate gradient methods with convergence rate equal to (in the case  $\tilde{\delta} = 0$  – Devolder–Glineur–Nesterov, 2013)

$$\mathcal{O}(1)\frac{LR^2}{N^p} + \mathcal{O}(1)N^{1-p}\tilde{\delta} + \mathcal{O}(1)N^{p-1}\delta,$$

where  $p \in [1, 2]$  (D. Kamzolov, 2017).

Assume that  $F$  is a **smooth** convex function and the gradient of  $F$  is Lipschitz-continuous with parameter  $L$ , then

$$0 \leq F(x) - F(y) - \langle \nabla F(y), x - y \rangle \leq \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in Q.$$

We can take  $\psi_{\delta_k}(x, y) = \langle \nabla F(y), x - y \rangle$ ,  $F_{\delta_k}(y) = F(y)$ , and  $\delta_k = 0$  for all  $k \geq 0$ . Let us assume that we can find exact solutions in intermediate optimization problems thus  $\tilde{\delta}_k = 0$  for all  $k \geq 0$  and

$$F(x_N) - F(x_*) \leq \frac{8LR^2}{(N+1)^2}.$$

The bound is optimal.

Let

$$\|\nabla F(x) - \nabla F(y)\|_* \leq L_\nu \|x - y\|^\nu, \quad \forall x, y \in Q.$$

Therefore,

$$0 \leq F(x) - F(y) - \langle \nabla F(y), x - y \rangle \leq \frac{L(\delta)}{2} \|x - y\|^2 + \delta, \quad \forall x, y \in Q,$$

where

$$L(\delta) = L_\nu \left[ \frac{L_\nu}{2\delta} \frac{1 - \nu}{1 + \nu} \right]^{\frac{1-\nu}{1+\nu}}$$

and  $\delta > 0$  is a parameter.

Consequently, we can take  $\psi_{\delta_k}(x, y) = \langle \nabla F(y), x - y \rangle$ ,

$F_{\delta_k}(y) = F(y)$ . Let us assume that  $\tilde{\delta}_k = 0$  for all  $k \geq 0$ . Let us take

$$\delta_k = \epsilon \frac{\alpha_{k+1}}{4A_{k+1}}, \quad \forall k,$$

where  $\epsilon$  is the accuracy of the solution by function.

In order to have

$$F(x_N) - F(x_*) \leq \epsilon,$$

it is enough to do

$$N \leq \inf_{\nu \in [0,1]} \left[ 64^{\frac{1+\nu}{1+3\nu}} \left( \frac{2-2\nu}{1+\nu} \right)^{\frac{1-\nu}{1+3\nu}} \left( \frac{L_\nu R^{1+\nu}}{\epsilon} \right)^{\frac{2}{1+3\nu}} \right]$$

steps.

The bound is optimal.

# Universal conditional gradient (Frank–Wolfe) method

In fast gradient method we have

$$\phi_{k+1}(x) = V(x, u_k) + \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}).$$

Instead of it let us take (idea goes back to A. Nemirovski, 2013)

$$\tilde{\phi}_{k+1}(x) = \alpha_{k+1}\psi_{\tilde{\delta}_k}(x, y_{k+1}).$$

Let us look at this substitution from the view of an error  $\tilde{\delta}_k$ . We can show that it enough to take  $\tilde{\delta}_k = 2R_Q^2$  for all  $k \geq 0$ , where  $R_Q$  is a diameter of a set  $Q$ . Let us take

$$\delta_k = \epsilon \frac{\alpha_{k+1}}{4A_{k+1}}, \forall k,$$

where  $\epsilon$  is the accuracy of the solution by function.

Finally,

$$N \leq \inf_{\nu \in (0,1]} \left[ 96^{\frac{1+\nu}{2\nu}} \left[ \frac{2-2\nu}{1+\nu} \right]^{\frac{1-\nu}{2\nu}} \left( \frac{L_\nu R_Q^{1+\nu}}{\epsilon} \right)^{\frac{1}{\nu}} \right].$$

$$F(x) \stackrel{\text{def}}{=} f(x) + h(x) \rightarrow \min_{x \in Q},$$

where  $f$  is a smooth convex function and the gradient of  $f$  is Lipschitz-continuous with parameter  $L$ . Function  $h$  is a convex function. We can show

$$\begin{aligned} 0 \leq F(x) - F(y) - \langle \nabla f(y), x - y \rangle - h(x) + h(y) &\leq \\ &\leq \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in Q. \end{aligned}$$

We can take  $\psi_{\delta_k}(x, y) = \langle \nabla f(y), x - y \rangle + h(x) - h(y)$ ,  $F_{\delta_k}(y) = F(y)$ , and  $\delta_k = 0$  for all  $k \geq 0$ . On the one hand the methods do not change, but on the other hand auxiliary optimizations are more complicated.

$$F(x) \stackrel{\text{def}}{=} f(f_1(x), \dots, f_m(x)) \rightarrow \min_{x \in Q},$$

where  $f_k$  is a smooth convex function and the gradient of  $f_k$  is Lipschitz-continuous with parameter  $L_k$  for all  $k \geq 0$ . Function  $f$  is a convex function, Lipschitz-continuous with parameter  $M$  with respect to  $L^1$ -norm, and monotonic function. Therefore,

$$0 \leq F(x) - F(y) - f(f_1(y) + \langle \nabla f_1(y), x - y \rangle, \dots, f_m(y) + \langle \nabla f_m(y), x - y \rangle) + F(y) \leq M \frac{\sum_{i=1}^m L_i}{2} \|x - y\|^2, \forall x, y \in Q.$$

We can take  $\psi_{\delta_k}(x, y) = f(f_1(y) + \langle \nabla f_1(y), x - y \rangle, \dots, f_m(y) + \langle \nabla f_m(y), x - y \rangle) - F(y)$ ,  $F_{\delta_k}(y) = F(y)$ , and  $\delta_k = 0$  for all  $k \geq 0$ .



# Proximal method with inexact solution in an intermediate optimization problem; Devolder–Glineur–Nesterov, 2013

Let us consider

$$f(x) \stackrel{\text{def}}{=} \min_{y \in Q} \underbrace{\left\{ \phi(y) + \frac{L}{2} \|y - x\|_2^2 \right\}}_{\Lambda(x,y)} \rightarrow \min_{x \in \mathbb{R}^n}. \quad (1)$$

Function  $\phi$  is a convex function and

$$\max_{y \in Q} \left\{ \Lambda(x, y(x)) - \Lambda(x, y) + \frac{L}{2} \|y - y(x)\|_2^2 \right\} \leq \delta.$$

Then

$$\left( \phi(y(x)) + \frac{L}{2} \|y(x) - x\|_2^2 - \delta, \psi_\delta(z, x) = \langle L(x - y(x)), z - x \rangle \right)$$

is  $(\delta, L)$ -model of  $f(z)$  at the point  $x$  w.r.t 2-norm. From this result one can obtain **Catalyst** technique (Lin–Marial–Harchaoui, 2015).

Let us consider

$$f(x) \stackrel{\text{def}}{=} \max_{y \in Q} [\langle x, b - Ay \rangle - \phi(y)] \rightarrow \min_{x \in \mathbb{R}^n}, \quad (2)$$

where  $\phi(y)$  is a  $\mu$ -strong convex function w.r.t.  $p$ -norm ( $1 \leq p \leq 2$ ). Then  $f$  is a smooth convex function and the gradient of  $f$  is Lipschitz-continuous with parameter

$$L = \frac{1}{\mu} \max_{\|y\|_p \leq 1} \|Ay\|_2^2.$$

If  $y_\delta(x)$  is a  $\delta$ -solution of the max problem, then

$$(\langle x, b - Ay_\delta(x) \rangle - \phi(y_\delta(x)), \psi_\delta(z, x) = \langle b - Ay_\delta(x), z - x \rangle)$$

is  $(\delta, 2L)$ -model of  $f(z)$  at the point  $x$  w.r.t 2-norm.

Let us consider

$$\phi(y) + \frac{\mu}{2} \|Ay - b\|_2^2 \rightarrow \min_{Ay=b, y \in Q}.$$

and it's dual problem

$$f(x) \stackrel{\text{def}}{=} \max_{y \in Q} \underbrace{\left( \langle x, b - Ay \rangle - \phi(y) - \frac{\mu}{2} \|Ay - b\|_2^2 \right)}_{\Lambda(x,y)} \rightarrow \min_{x \in \mathbb{R}^n}.$$

If  $y_\delta(x)$  is a solution in the sense of an inequality

$$\max_{y \in Q} \langle \nabla_y \Lambda(y_\delta(x), y), y - y_\delta(x) \rangle \leq \delta,$$

then

$$\begin{aligned} & \left( \langle x, b - Ay_\delta(x) \rangle - \phi(y_\delta(x)) - \frac{\mu}{2} \|Ay_\delta(x) - b\|_2^2, \right. \\ & \left. \psi_\delta(z, x) = \langle b - Ay_\delta(x), z - x \rangle \right) \end{aligned}$$

is  $(\delta, \mu^{-1})$ -model of  $f(z)$  at the point  $x$  w.r.t 2-norm.

# First new example: Min-min problem

Let us consider

$$f(x) \stackrel{\text{def}}{=} \min_{y \in Q} F(y, x) \rightarrow \min_{x \in \mathbb{R}^n} .$$

- Set  $Q$  is convex and bounded.
- Function  $F$  is smooth and convex w.r.t. both variables.
- $\|\nabla F(y', x') - \nabla F(y, x)\|_2 \leq L \|(y', x') - (y, x)\|_2, \forall y, y' \in Q, x, x' \in \mathbb{R}^n$ .

If we can find a point  $\tilde{y}_\delta(x) \in Q$  such that

$$\langle \nabla_y F(\tilde{y}_\delta(x), x), y - \tilde{y}_\delta(x) \rangle \geq -\delta, \quad \forall y \in Q,$$

then

$$F(\tilde{y}_\delta(x), x) - f(x) \leq \delta, \quad \|\nabla f(x') - \nabla f(x)\|_2 \leq L \|x' - x\|_2$$

and

$$(F(\tilde{y}_\delta(x), x) - 2\delta, \psi_\delta(z, x) = \langle \nabla_y F(\tilde{y}_\delta(x), x), z - x \rangle)$$

is  $(6\delta, 2L)$ -model of  $f(x)$  at the point  $x$  w.r.t 2-norm.

## Second new example: Nesterov's clustering model, 2018

Let us consider

$$f_\mu(x = (z, p)) = g(z, p) + \mu \sum_{k=1}^n z_k \ln z_k + \frac{\mu}{2} \|p\|_2^2 \rightarrow \min_{z \in S_n(1), p \geq 0}.$$

Let's introduce norm

$$\|x\|^2 = \|(z, p)\|^2 = \|z\|_1^2 + \|p\|_2^2.$$

Assume that

$$\|\nabla g(x_2) - \nabla g(x_1)\|_* \leq L \|x_2 - x_1\|, \quad L \leq \mu.$$

Then

$$\left( f_\mu(y), \langle \nabla g(y), x - y \rangle + (\mu - L) \sum_{k=1}^n z_k \ln z_k + \frac{\mu - L}{2} \|p\|_2^2 \right).$$

is  $(0, 2L)$ -model of  $f_\mu(x = (z, p))$  at the point  $y$ .

## Third new example: Proximal Sinkhorn algorithm

Let introduce smoothed Wasserstein distance (Peyre–Cuturi, 2018)

$$W_\gamma(p, q) = \min_{\pi \in \Pi(p, q)} \left\{ \sum_{i, j=1}^n C_{ij} \pi_{ij} + \gamma \sum_{i, j} \pi_{ij} \log \frac{\pi_{ij}}{\pi_{ij}^0} \right\},$$

where

$$\Pi(p, q) = \left\{ \pi \in \mathbb{R}_+^{n \times n}, \sum_{j=1}^n \pi_{ij} = p_i, \sum_{i=1}^n \pi_{ij} = q_j \right\}.$$

It's well known (Franklin–Lorentz, 1989; Dvurechensky et al., 2018) that the complexity (a.o.) of the Sinkhorn algorithm for this problem is

$$O \left( n^2 \min \left\{ \frac{\ln n}{\gamma \epsilon}, \exp \left( \frac{\tilde{C}}{\gamma} \right) \ln(\epsilon^{-1}) \right\} \right).$$

## Third new example: Proximal Sinkhorn algorithm

If we want to calculate  $W_0(p, q)$  with the precision  $\epsilon$  one can calculate  $W_\gamma(p, q)$  with  $\gamma = \epsilon/4 \ln n$  with the precision  $\epsilon/2$ . So the complexity in this case will be  $\tilde{O}(n^2/\epsilon^2)$  a.o. Can we do better?

**The answer is Yes!.**

If we use **proximal model** of  $f(x)$  at point  $y$  of the form

$$\psi_\delta(x, y) = f(x) - f(y)$$

and choose arbitrary  $L > 0$ . Then proximal gradient descent (Chen–Teboulle, 1993)

$$x^{k+1} = \arg \min_{x \in Q} \left\{ \psi_\delta(x, x^k) + LV(x, x^k) \right\}$$

will converges as  $O(LR^2/k)$ ,  $R^2 = V(x_*, x^0)$ .

## Third new example: Proximal Sinkhorn algorithm

So, let's apply proximal set up to  $W_0(p, q)$  calculation. Put  $x = \pi$ ,  $Q = \Pi$ ,  $L = \gamma$ , choose  $V = KL$ . We obtain that proximal gradient method

$$\pi^{k+1} = \min_{\pi \in \Pi(p, q)} \left\{ \sum_{i, j=1}^n C_{ij} \pi_{ij} + \gamma \sum_{i, j} \pi_{ij} \log \frac{\pi_{ij}}{\pi_{ij}^k} \right\}$$

requires  $O(\gamma \ln n / \epsilon)$  iterations to reach the precision  $\epsilon$ . But, this assumes that we have to solve exactly auxiliary problem – i.e. we can calculate exactly  $W_\gamma(p, q)$ . But from above we know that if  $\gamma$  is not small (**Note:** we can choose  $\gamma$  as we want!) we can solve this problem with very high precision with the complexity  $\tilde{O}(n^2)$ . So the total number of a.o. will be  $\tilde{O}(n^2 / \epsilon)$  – that is better than for the simple Sinkhorn  $\tilde{O}(n^2 / \epsilon^2)$ .



## Auxiliary optimization problem

Let us take  $\phi_{k+1}(x) = V(x, u_k) + \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1})$ . From

$$\langle \nabla \phi_{k+1}(x_\epsilon), x_\epsilon - x_* \rangle \leq \epsilon$$

we can get

$$\phi_{k+1}(x_\epsilon) - \phi_{k+1}(x_*) \leq \epsilon.$$

Let an intermediate optimization problem has Lipschitz-continuous with parameter  $M$  function. Let us assume  $x_\epsilon$  is a  $\epsilon$ -solution then

$$\langle \nabla \phi_{k+1}(x_\epsilon), x_\epsilon - x_* \rangle \leq \|\nabla \phi_{k+1}(x_\epsilon)\|_* \|x_\epsilon - x_*\| \leq M\sqrt{2\epsilon}.$$

If we have linear convergence speed for an auxiliary optimization problem, then we should do  $c$  times more steps where  $c$  is a small constant.

Thank you!