Motivation

- Consider the minimization of the composite finite-average of many functions:
  \[ \min_x \phi(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) + h(x), \]
  where \( f_i \) are twice continuously differentiable and convex, \( h \) is closed convex.
- Big data setting: \( n \) is very large (millions, billions etc.).
- Incremental/stochastic optimization methods, which process only one \( f_i \) at each iteration, are among the most effective methods for this task.
- There exists many different incremental optimization schemes:
  - SGD, dLBFGS [Schraudolph et al., 2007], AdaGrad [Duchi et al., 2011], SQN [Byrd et al., 2014], Adam [Kingma, 2014] etc.
  - SAC [Schmidt et al., 2015], SWRG [Johnson & Zhang, 2013], SAGA [Defazio et al., 2014a], MISD [Mairal, 2015].
- They all have either a sublinear or linear convergence rate.
- Goal: an incremental optimization method with a superlinear rate of convergence.

Main idea

- Build the second-order Taylor approximation of each \( f_i \):
  \[ m_k^i(x) := f_i(v_k^i) + \nabla f_i(v_k^i)^	op (x - v_k^i) + \frac{1}{2} (x - v_k^i)^	op \nabla^2 f_i(v_k^i) (x - v_k^i). \]
- Then \( \phi \) can be approximated with \( m_k(x) := \frac{1}{n} \sum_{i=1}^{n} m_k^i(x) + h(x). \)
- Find the minimizer of the model: \( \tilde{x}_k := \arg\min_x m_k(x) \)
- Choose next iterate \( x_{k+1} \) between \( x_k \) and \( \tilde{x}_k \): \( x_{k+1} := x_k + \alpha_k (\tilde{x}_k - x_k). \)
- Each time update only one \( v_k^i \) to keep the iteration cost independent of \( n \).

In general, there is no need to find the minimizer \( \tilde{x}_k \) of the model exactly.

Define the composite gradient mapping:
\[
T_k(x, \xi) := \arg\min_y \left\{ y^\top y + \frac{L}{2} \| y - x \|^2 + h(y) \right\},
\]
where \( L \geq 0 \) we have \( G_k(x, \xi) \equiv \xi \).

We show that, instead of \( \tilde{x}_k := \arg\min_x [m_k(x) =: s_k(x) + h(x)] \), any point \( \tilde{x}_k := T_k(x; \xi) \) can be used in NIM provided that
\[
\| G_k(x; \tilde{x}_k) \| \leq \min \{ (\Delta_k^2)^\gamma, \Delta_k \}, \quad \Delta_k := \| G_k(\xi, s_k) \|,
\]
Here \( L \) can be any such that \( L \geq L_0 \equiv 1, \tilde{v}_k := \frac{1}{n} \sum_{i=1}^{n} v_i, \) and \( \gamma \in (0, 1] \).

Intuition: the closer NIM to the optimum, the more accurate \( s_k \) is required.

Possible inner solver: Fast Gradient Method [Nesterov, 2013].