none/global//global/global 0Nesterov18 00Nesterov18 0Nesterov18 00Nesterov18 0ErnVohralik13a 00ErnVohralik13a 0HeidWihler20 00HeidWihler20 0HeidPraetoriusWihler21 00HeidPraetoriusWihler21 0HaberlPraetoriusSchimanko21 00HaberlPraetoriusSchimanko21 0GantnerHaberlPraetorius21 00GantnerHaberlPraetorius21 0HeidWihler21 00HeidWihler21 0HeidStammWihler21 00HeidStammWihler21 0HeidWihler22 00HeidWihler22 5AB31E7DB2443CAE55F3314E362E1448 0Nesterov18none/global//global/gl
0ErnVohralik13anone/global//global/global 0HeidWihler20none/global//global/global
0HeidPraetoriusWihler21none/global//global/global 0HaberlPraetoriusSchimanko21none/global//global/global 0GantnerHaberlPraetorius21none/global//global/gl
0HeidWihler21none/global//global/global 0HeidStammWihler21none/global//global/globa
0HeidWihler22none/global//global/global 0Nesterov18 0ErnVohralik13a 0HeidWihler20 0HeidPraetoriusWihler21 0HaberlPraetoriusSchimanko21 0GantnerHaberlPraetorius21 0HeidWihler21 0HeidStammWihler21 0HeidWihler22

Optimization algorithms with approximation

December 11, 2022

Contents

1 Introduction

We consider a Hilbert space $(X, \langle\cdot,\cdot\rangle)$ with induced norm $\|\cdot\|$ and the minimization of a smooth μ-strictly convex function $f : X \to \mathbb{R}$:

$$\inf_{x\in X} f(x) = \inf\{f(x) \mid x \in X\}.$$

We suppose that a unique minimizer x^* exists.

Our purpose is to analyse gradient algorithms on a sequence of subspaces (finite element spaces for the PDE)

$$X_0 \subset \cdots \subset X_k \subset X_{k+1} \subset \cdots \subset X, \quad P_k : X \to X_k,$$

such that a typical iteration reads:

$$x_{k+1} = x_k - t_k P_k \nabla f(x_k), \tag{1}$$

where P_k is the orthogonal projector on X_k and $\nabla f(x) \in X$ is defined by the Riesz map. In order to generate the subspaces X_k, we suppose to have an error estimators $\eta_k : X_k \to \mathbb{R}$ and a refinement algorithm satisfying typical hypothesis from the theory of AFEM. In the case $f \in \mathcal{S}^{1,1}_{\mu,L}(X)$, X finite-dimensional, and $t_k = 2/(\mu + L)$ for all k we have the following convergence estimate (Theorem 2.1.15 in [**?**]) for the gradient method (**GM**):

$$\|x_n - x^*\| \leqslant \rho^n \|x_0 - x^*\|, \quad \rho = 1 - 1/\kappa, \tag{2}$$

such that $\varepsilon > 0$ is achieved in $n(\varepsilon) = O(\kappa)\ln(1/\varepsilon)$ iterations. It is well-known that **GM** is not optimal for this class of functions. The accelerated gradient method (**AGM**) [**?**] yields an improved estimate $n(\varepsilon) = O(\sqrt{\kappa})\ln(1/\varepsilon)$.

Our aim is to establish a similar iteration count for the method on a sequence of subspaces. There is important progress of adaptive finite element methods (AFEM) for nonlinear elliptic equations, see [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**], and our development is based on these works. However, here, we wish to work out the optimization point of view.

2 Notation

We throughout suppose that $f : X \to \mathbb{R}$ is convex and C^1 and we use the the Fréchet-Riesz theorem to define

$$\langle \nabla f(y), x \rangle = f'(y)(x) \quad \forall x, y \in X.$$

It is then easy to see, that for all y in a closed subspace $Y \subset X$ and $P_Y : X \to Y$ its orthogonal projector

$$P_Y \nabla f(y) = \nabla f_{|Y}. \tag{3}$$

3 Assumptions on subspace selection

Let $X_0 \subset X$ be a subspace. We suppose to have a lattice of admissible closed subspaces

$$\mathcal{X}(X_0) = \{X_0 \subset Y \subset X\}. \tag{4}$$

The partial order on $\mathcal{X}(X_0)$ is given by $Y \geqslant Z$ if and only if Y is a superspace of Z. We then have the finest common coarsening $Y \wedge Z$ and the coarsest common refinement $Y \vee Z$, respectively. We let

$$\mathcal{X}(Y) = \{Z \in \mathcal{X}(X_0) \mid Y \wedge Z = Y\}, \quad Y \in \mathcal{X}(x_0).$$

We make the following assumptions. First we have a reliable error estimator η such that for all $Y \in \mathcal{X}(X_0)$

$$\|(I - P_Y)\nabla f(y)\| \leqslant C_{\text{rel}}\eta(y, Y) \qquad \forall y \in Y \tag{H1}$$
$$|\eta(y, Y) - \eta(z, Y)| \leqslant C_{\text{stab}} \|y - z\| \qquad \forall y, z \in Y \tag{H2}$$

and a subspace generator $Y^+ = \mathbf{GSp}(Y, y)$ such that with $0 \leqslant q_{\text{red}} < 1$

$$\eta^2(y, Y^+) \leqslant q_{\text{red}}\eta^2(y, Y) \qquad \forall y \in Y \tag{H2}$$

4 Gradient method with constant step-size

Here we suppose in addition that f has a L-Lipschitz continuous gradient

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L \|x - y\| \quad \forall x, y \in X.$$

Setting $\beta = 0$ in the following algorithm, we have the standard gradient method with fixed step-size.

Algorithm 1: GM with constant step-size

Inputs: $X_0, x_0 \in X_0$, $t_0 > 0, 0 \leqslant \beta < 1, \lambda > 0$. Set $x_{-1} = x_0$ and $k = 0$.

(1) $y_k = (1 + \beta)x_k - \beta x_{k-1}$.

(2) $x_{k+1} = y_k - \frac{1}{L} P_{X_k} \nabla f(y_k)$.

(3) If $\eta(x_{k+1}, X_k) \geqslant q_{\text{red}}\eta(x_k, X_k) + \lambda(f(x_k) - f(x_{k+1})$:
$X_{k+1} = \mathbf{GSp}(X_K, x_{k+1})$,
Else: $X_{k+1} = X_k$.

(4) Increment k and go to (1).

We will write for brevity $P_k = P_{X_k}$ etc. By the Lipschitz-condition and convexity we have for any $x \in X$ with (3)

$$\begin{aligned}
f(x_{k+1}) \leqslant & f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{L}{2} \|P_k \nabla f(y_k)\|^2 \\
= & f(y_k) + \langle P_k \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{L}{2} \|P_k \nabla f(y_k)\|^2 \\
\leqslant & f(x) + \langle \nabla f(y_k), y_k - x \rangle - \frac{\mu}{2} \|x - y_k\|^2 - \frac{1}{2L} \|P_k \nabla f(y_k)\|^2
\end{aligned}$$

Let $\theta = \frac{1-\beta}{1+\beta}$. Taking the last inequality θ-times with $x = x^*$ and $1-\theta$-times with $x = x_k$, we have, setting $\Delta f_k := f(x_k) - f(x^*)$

$$\begin{aligned}
\Delta f_{k+1} - (1-\theta)\Delta f_k \leqslant & \langle \nabla f(y_k), y_k - (1-\theta)x_k - \theta x^* \rangle - \frac{1}{2L} \|P_k \nabla f(y_k)\|^2 \\
& - \frac{\mu\theta}{2} \|x^* - y_k\|^2 - \frac{\mu(1-\theta)}{2} \|x_k - y_k\|^2
\end{aligned}$$

Let

$$R_k := \theta \langle (I - P_k) \nabla f(y_k), y_k - x^* \rangle,$$

such that

$$\langle \nabla f(y_k), y_k - (1-\theta)x_k - \theta x^* \rangle = R_k + \langle P_k \nabla f(y_k), y_k - (1-\theta)x_k - \theta x^* \rangle$$

and

$$z_k := \frac{x_k}{\theta} - \frac{1-\theta}{\theta} x_{k-1} = x_k + \frac{1-\theta}{\theta}(x_k - x_{k-1}).$$

We also have with $\theta(1+\beta) = 1 - \beta$

$$z_k = y_k + \frac{1-\theta-\theta\beta}{\theta\beta}(y_k - x_k) = y_k + \frac{y_k - x_k}{\theta}$$

Then with $2ab - a^2 = b^2 - (a-b)^2$

$$\begin{aligned}
& \langle P_k \nabla f(y_k), y_k - (1-\theta)x_k - \theta x^* \rangle - \frac{1}{2L} \|P_k \nabla f(y_k)\|^2 = \\
& \frac{L}{2} \left(\|y_k - (1-\theta)x_k - \theta x^*\|^2 - \|x_{k+1} - (1-\theta)x_k - \theta x^*\|^2 \right) = \\
& \frac{\theta^2 L}{2} \left(\left\| x_k + \frac{y_k - x_k}{\theta} - x^* \right\|^2 - \|z_{k+1} - x^*\|^2 \right) = \frac{\theta^2 L}{2} \left(\|z_k - (y_k - x_k) - x^*\|^2 - \|z_{k+1} - x^*\|^2 \right)
\end{aligned}$$

Since with $-2ab = (a-b)^2 - a^2 - b^2$

$$
\begin{aligned}
\|z_k - (y_k - x_k) - x^*\|^2 &= \|z_k - x^*\|^2 - 2\theta\langle z_k - x^*, z_k - y_k\rangle + \|y_k - x_k\|^2 \\
&= (1-\theta)\|z_k - x^*\|^2 + (\theta^2 - \theta)\|z_k - y_k\|^2 + \theta\|y_k - x^*\|^2
\end{aligned}
$$

we have

$$
\Delta f_{k+1} + (1-\theta)\Delta f_k \leqslant \frac{\theta^2 L}{2}\left((1-\theta)\|z_k - x^*\|^2 - \|z_{k+1} - x^*\|^2\right) + \left(\frac{\theta^3 L}{2} - \frac{\theta\mu}{2}\right)\|y_k - x^*\|^2 + R_k
$$

We have

$$
\begin{aligned}
(1-\theta)\|z_k - x^*\|^2 + \theta\|y_k - x^*\|^2 &= \|(1-\theta)z_k + \theta y_k - x^*\|^2 + \theta(1-\theta)\|z_k - y_k\|^2 \\
&= \left\|y_k + (1-\theta)\frac{y_k - x_k}{\theta} - x^*\right\|^2 + \frac{1-\theta}{\theta}\|y_k - x_k\|^2
\end{aligned}
$$

$$
R_k \leqslant \frac{1}{\mu}\|(I - P_k)\nabla f(y_k)\|^2 + \frac{\theta\mu}{4}\|y_k - x^*\|^2 \leqslant \frac{C_{\mathrm{rel}}^2}{\mu}\eta^2(X_k, y_k) + \frac{\theta\mu}{4}\|y_k - x^*\|^2
$$

If the criterion in step (3) of the algorithm does not hold, we have

$$
\eta(x_{k+1}, X_k) \leqslant q_{\mathrm{red}}\eta(x_k, X_k) + \lambda(f(x_k) - f(x_{k+1})
$$

Otherwise, we have

$$
\eta(x_{k+1}, X_k) \leqslant q_{\mathrm{red}}\eta(x_{k+1}, X_k) \leqslant q_{\mathrm{red}}\eta(x_k, X_k) + C_{\mathrm{stab}}\|x_{k+1} - x_k\|
$$