New Methods for Regularization Path Optimization via Differential Equations

Heyuan Liu
Department of Industrial Engineering and Operations Research
University of California, Berkeley
Berkeley, CA 94720
heyuan_liu@berkeley.edu

Paul Grigas
Department of Industrial Engineering and Operations Research
University of California, Berkeley
Berkeley, CA 94720
pgrigas@berkeley.edu

Abstract

We develop and analyze several second order algorithms for computing an approximately optimal solution (regularization) path of a parameterized convex optimization problem with smooth Hessian. Our algorithms are inspired by a differential equations perspective on the parametric solution path and do not rely on the specific structure of the regularizer. We present computational guarantees that bound the oracle complexity to achieve an approximately optimal solution path under different smoothness assumptions and that also hold in the presence of approximate subproblems. We conduct numerical experiments that demonstrate the viability of our approach, especially in the presence of higher-order smoothness.

1 Introduction

We consider the parametric optimization problem

$$P(\lambda): \quad F_\lambda^* := \min_{x \in \mathbb{R}^p} \left\{ F_\lambda(x) := f(x) + \lambda \cdot \Omega(x) \right\}, \tag{1}$$

where $f(\cdot) : \mathbb{R}^p \to \mathbb{R}$ and $\Omega(\cdot) : \mathbb{R}^p \to \mathbb{R}$ are twice-differentiable functions such that $f(\cdot)$ is μ-strongly convex for some $\mu \geq 0$ and $\Omega(\cdot)$ is 1-strongly convex, both with respect to the ℓ_2-norm (denoted by $\|\cdot\|$ herein). For any $\lambda > 0$, let $x(\lambda) := \arg\min_{x \in \mathbb{R}^p} \{F_\lambda(x)\}$ denote the unique optimal solution of $P(\lambda)$ defined in (1). We are interested in the problem of computing the set of optimal solutions $\{x(\lambda) : \lambda \in [\lambda_{\min}, \lambda_{\max}]\}$ for some $0 < \lambda_{\min} < \lambda_{\max}$, and we also refer to this set of solutions as the *(exact) regularization path*. Indeed, in machine learning applications, x may represent a vector of model parameters, $f(\cdot)$ may represent a loss function to measure the fit or performance of the model on the training data, and $\Omega(\cdot)$ may represent a regularizer that is intended to mitigate the danger of overfitting. In such applications, one needs to solve $P(\lambda)$ for many different values of λ, using cross validation techniques to tune the final value of λ. More generally, the parametric optimization problem (1) and our algorithms allow for $\Omega(\cdot)$ to be more complicated than a simple regularization function, as in the case of re-weighted logistic regression examined in Section 4.

Let us describe a differential equations perspective on the regularization path that will prove fruitful in developing efficient computational methods. First we introduce a re-parametrization in terms of an auxiliary variable $t \geq 0$ (thought of as "time"), whereby for a given $T > 0$ we introduce

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

functions $\lambda(\cdot) : [0, T] \to [\lambda_{\min}, \lambda_{\max}]$ and $\xi(\cdot) : [\lambda_{\min}, \lambda_{\max}] \to \mathbb{R}$ such that $\xi(\cdot)$ is Lipschitz, $\lambda(\cdot)$ is differentiable on $(0, T)$, and it holds that $\frac{\mathrm{d}\lambda}{\mathrm{d}t} = \xi(\lambda(t))$ for all $t \in (0, T)$. In a slight abuse of notation, we define the path with respect to t as $x(t) := x(\lambda(t))$. Now notice that, for any $t \in [0, T]$, the first-order optimality condition for problem $P(\lambda(t))$ states that $\nabla f(x(t)) + \lambda(t)\nabla\Omega(x(t)) = 0$. By differentiating both sides of the previous equation with respect to t, it holds that $\nabla^2 f(x(t)) \cdot \frac{\mathrm{d}x}{\mathrm{d}t} + \nabla\Omega(x(t)) \cdot \frac{\mathrm{d}\lambda}{\mathrm{d}t} + \lambda(t)\nabla^2\Omega(x(t)) \cdot \frac{\mathrm{d}x}{\mathrm{d}t} = 0$. Rearranging the above and again using $\nabla f(x(t)) + \lambda(t)\nabla\Omega(x(t)) = 0$ yields $\frac{\mathrm{d}x}{\mathrm{d}t} = \left(\nabla^2 f(x(t)) + \lambda(t)\nabla^2\Omega(x(t))\right)^{-1} \frac{\xi(\lambda(t))}{\lambda(t)} \nabla f(x(t))$. Thus, we arrive at the following autonomous system

$$\frac{\mathrm{d}\lambda}{\mathrm{d}t} = \xi(\lambda), \quad \frac{\mathrm{d}x}{\mathrm{d}t} = v(x, \lambda) := \left(\nabla^2 f(x) + \lambda\nabla^2\Omega(x)\right)^{-1} \frac{\xi(\lambda)}{\lambda} \nabla f(x), \text{ for } t \in [0, T]. \quad (2)$$

Contributions and perspective. The above derivation provides a novel and intuitive perspective of the regularization path from an ordinary differential equations point of view. Furthermore, the above does not rely on any special structural assumptions on $f(\cdot)$ or $\Omega(\cdot)$ – besides twice-differentiability of $f(\cdot)$ and $\Omega(\cdot)$ and strong convexity of $\Omega(\cdot)$ – whereas most existing approaches for computing the regularization path only apply to special cases such as the LASSO or SVM problems [20, 2, 21] (see, for example, [15, 4, 8, 18, 5]). The case of more general $\Omega(\cdot)$ functions has received relatively less attention and most existing works have focused on upper and lower complexity bounds using an equally spaced grid of λ values and/or piece-wise constant mappings [17, 7, 10].

The autonomous system (2) provides a starting point for our algorithms for computing the regularization path developed herein, whereby we will consider several different discretizations of the above system. Interestingly, the dynamics in (2) resemble the dynamics of Newton's method as do the algorithms we develop in Section 2. In contrast, our focus herein is on computing the *entire* solution path as opposed to a single optimization problem for a given (typically small) value of λ, and we do not make any self-concordance assumptions on the objective functions as is often done in the analysis of Newton-type methods such as interior point methods (see, for example, [9, 16]).

In Section 2, we develop algorithms based on combining semi-implicit Euler and trapezoidal discretization methods for (2) with interpolation schemes. In contrast to the typical asymptotic analysis of numerical methods for differential equations, we develop computational guarantees for these algorithms that directly bound the number of first and second order oracle calls and linear system solves required to generate an approximate solution path. In Section 3, we consider algorithms that allow for approximate subproblems and that only rely on gradient and Hessian-vector product computations, and in Section 4 we validate our methods on some logistic regression examples.

2 Algorithms and complexity results

In this section, we present computational algorithms for computing an approximate regularization path based on discretizations of (2), along with corresponding complexity analysis. The primary error metric that we consider is the ℓ_2 norm of the gradient across the entire interval $[\lambda_{\min}, \lambda_{\max}]$.

Definition 1. An approximate regularization path $\hat{x}(\cdot) : [\lambda_{\min}, \lambda_{\max}] \to \mathbb{R}^p$ to the parametric optimization problem (1) has accuracy $\epsilon \geq 0$ if $\|\nabla F_\lambda(\hat{x}(\lambda))\| \leq \epsilon$ for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$.

Notice that the strong convexity of the objective function $F_\lambda(\cdot)$ for all $\lambda > 0$ immediately implies that an ϵ-accurate regularization path $\hat{x}(\cdot)$ also satisfies $F_\lambda(\hat{x}(\lambda)) - F_\lambda^* \leq \frac{\epsilon^2}{2(\mu+\lambda)}$ for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. We develop oracle complexity results for our methods in terms of the number of gradient computations, Hessian computations, and linear system solves required to compute an ϵ-accurate approximate path. In our complexity analysis, we make the following smoothness assumptions on $f(\cdot)$ and $\Omega(\cdot)$.

Assumption 1. *In addition to μ-strong convexity of $f(\cdot)$ and 1-strong convexity of $\Omega(\cdot)$, these functions have L-Lipschitz gradients and Hessians, where $L > 0$ is an upper bound on the four relevant Lipschitz constants. In addition, we assume that $f^* := F_0^* > -\infty$, and that $G > 0$ is an upper bound on the norm of the gradients of $f(\cdot)$ and $\Omega(\cdot)$ on the level set $\{x \in \mathbb{R}^p : f(x) \leq f(x_0)\}$.*

Algorithm 1 below presents a two-step "meta-algorithm" for computing an approximate regularization path $\hat{x}(\cdot)$. Inspired by numerical methods to solve ordinary differential equations, we first design several schemes to iteratively update (x_k, λ_k) by exploiting the dynamics in (2). We use the function

$\psi(\cdot,\cdot) : \mathbb{R}^p \times [\lambda_{\min}, \lambda_{\max}] \to \mathbb{R}^p \times [\lambda_{\min}, \lambda_{\max}]$ to denote a generic update rule in the meta-algorithm below, and we consider several different specific examples herein. Then we apply an interpolation method $\mathcal{I}(\cdot)$ to resolve the previously computed sequence of points into an approximate path $\hat{x}(\cdot)$.

Algorithm 1: Meta-algorithm for computing an approximate regularization path $\hat{x}(\cdot)$

input : Initial point $x_0 \in \mathbb{R}^p$, initial regularization parameter $\lambda_0 \leftarrow \lambda_{\max}$, total number of iterations $K \geq 1$, update rule $\psi(\cdot,\cdot)$, and interpolation method $\mathcal{I}(\cdot)$.

for $k = 0, \ldots, K-1$ **do**

  Update $(x_{k+1}, \lambda_{k+1}) \leftarrow \psi(x_k, \lambda_k)$

end

Output $\hat{x}(\cdot) \leftarrow \mathcal{I}\left(\{(x_k, \lambda_k)\}_{k=1}^K\right)$

Euler update scheme. Let us now present the simplest version of Algorithm 1 based on applying Euler's method to specify the update rule $\psi(\cdot,\cdot)$. For simplicity, from now on we only consider the case where $\xi(\lambda) = -\lambda$ in (2), although our algorithms and complexity analysis also hold for a broad family of functions $\xi(\cdot)$, or equivalently $\lambda(\cdot)$. In this case, the Euler update rule $\psi_{\mathrm{E}}(\cdot,\cdot)$ is given by

$$\psi_{\mathrm{E}}(x_k, \lambda_k) := \left((1-h)\lambda_k, x_k - hv(x_k, (1-h)\lambda_k)\right), \tag{3}$$

where $h := 1 - (\frac{\lambda_{\min}}{\lambda_{\max}})^{1/K}$ is a step-size parameter depending on the total number of iterations $K \geq 1$. Notice that the update of x uses $(1-h)\lambda_k = \lambda_{k+1}$ instead of λ_k, which can be regarded as a semi-implicit Euler update. One simple way to implement the interpolation method $\mathcal{I}(\cdot)$ is with linear interpolation, whereby $\hat{x}(\lambda) := \alpha x_k + (1-\alpha)x_{k+1}$ with $\alpha = \frac{\lambda - \lambda_{k+1}}{\lambda_k - \lambda_{k+1}}$ for all $\lambda \in [\lambda_{k+1}, \lambda_k]$ and $k \in \{0, \ldots, K-1\}$. Theorem 1 is our main result concerning the complexity of Algorithm 1 with update scheme (3) and linear interpolation and demonstrates that, in terms of the accuracy parameter ϵ, the Euler update rule requires $O(1/\epsilon)$ iterations to compute a 4ϵ-accurate regularization path.

Theorem 1. *Suppose that Assumption 1 holds, let $\epsilon > 0$ be the desired accuracy, and suppose that the initial point x_0 satisfies $\|\nabla F_{\lambda_0}(x_0)\| \leq \epsilon$. Let $T := \log(\lambda_{\max}/\lambda_{\min})$, let $\tilde{\mu} := \mu + \lambda_{\min}$, and let*

$$K_{\mathrm{E}} := \left\lceil \max\left\{ 2T, \frac{\sqrt{LG}T}{\sqrt{3}\tilde{\mu}}, \frac{(f(x_0) - f^*)L(1+\tilde{\mu})T}{\epsilon\tilde{\mu}}, \frac{\sqrt{L}(G+1)(1+\tilde{\mu})T}{\sqrt{\epsilon}\tilde{\mu}} \right\} \right\rceil.$$

If the total number of iterations K satisfies $K \geq K_{\mathrm{E}}$, then Algorithm 1 with the Euler update rule (3) *and linear interpolation scheme returns a 4ϵ-accurate regularization path.*

Trapezoid update scheme. Motivated by multi-stage numerical methods for solving differential equations, we also propose a version of Algorithm 1 that uses the trapezoid update scheme, denoted by $\psi_{\mathrm{tr}}(\cdot,\cdot)$ and where $(x_{k+1}, \lambda_{k+1}) = \psi_{\mathrm{tr}}(x_k, \lambda_k)$ is defined by the recursive process in (4) below.

$$\begin{array}{lll} x_{k,1} = x_k, & \lambda_{k,1} = \lambda_k, & d_{k,1} = v(x_{k,1}, \lambda_{k,1}), \\ x_{k,2} = x_k + hd_{k,1}, & \lambda_{k,2} = (1 - h + h^2)\lambda_k, & d_{k,2} = v(x_{k,2}, \lambda_{k,2}), \\ d_k = \frac{d_{k,1} + d_{k,2}}{2}, & \lambda_{k+1} = (1 - h + \frac{h^2}{2})\lambda_k, & x_{k+1} = x_k + hd_k. \end{array} \tag{4}$$

The trapezoid update is beneficial when $f(\cdot)$ and $\Omega(\cdot)$ have higher-order smoothness properties and ϵ is relatively small, and Theorem 2 below verifies that using the trapezoid update rule in Algorithm 1 leads to an improved $O(1/\sqrt{\epsilon})$ complexity under higher order smoothness. One may also consider applying an update scheme that takes advantage of even higher-order smoothness properties such as the Runge-Kutta update rule [1] with cubic interpolation [6], which we consider in our experiments.

Theorem 2. *Suppose that Assumption 1 holds and that $f(\cdot)$ and $\Omega(\cdot)$ additionally have L-Lipschitz third order directional derivatives. Let $\epsilon > 0$ be the desired accuracy, let $\tilde{\mu} := \mu + \lambda_{\min}$, suppose that the initial point x_0 satisfies $\|\nabla F_{\lambda_0}(x_0)\| \leq \epsilon \leq \tilde{\mu}$, let $T := 1.1\log(\lambda_{\max}/\lambda_{\min})$, and let*

$$K_{\mathrm{tr}} := \left\lceil \max\left\{ 10T, \frac{8LT(1+G)}{\tilde{\mu}}, \frac{6L^{1/2}(1+G)^{3/2}T}{\epsilon^{1/2}}, \frac{5L(1+G)^{4/3}T}{\tilde{\mu}^{2/3}\epsilon^{1/3}} \right\} \right\rceil.$$

If the total number of iterations K satisfies $K \geq K_{\mathrm{tr}}$, then Algorithm 1 with the trapezoid update rule (4) *and linear interpolation scheme returns a 2ϵ-accurate regularization path.*

3 Approximate subproblem and second-order conjugate gradient extensions

Each iteration of both the Euler and trapezoid update methods involves computing Hessian matrices and solving a system of linear equations in order to obtain $v(x,\lambda) = (\nabla^2 f(x) + \lambda \nabla^2 \Omega(x))^{-1} \nabla f(x)$. In large scale problems, these computations may become burdensome. In this section we present extensions of our previous complexity results for the Euler and trapezoid method that hold in the presence of approximate computation of $v(\cdot,\cdot)$. Formally, let us define $\hat{d} \in \mathbb{R}^p$ to be δ-accurate approximation of $v(x,\lambda)$ if $\|(\nabla^2 f(x) + \lambda \nabla^2 \Omega(x))\hat{d} - \nabla f(x)\| \leq \delta$. The following proposition summarizes our results when replacing $v(x,\lambda)$ with a δ-accurate approximation.

Proposition 1. *For a given desired accuracy $\epsilon > 0$, consider modifying the previously examined Euler and trapezoid variants of Algorithm 1 by replacing all computations of the direction $v(\cdot,\cdot)$ with a δ-accurate approximation for some $\delta \in [0,\epsilon]$. Then, under the same conditions as Theorem 1, if the total number of iterations K satisfies $K \geq K_{\mathrm{E}}$, then the δ-approximate variant of Algorithm 1 with the Euler update rule* (3) *returns a 5ϵ-accurate regularization path. Likewise, under the same conditions as Theorem 2, if the total number of iterations K satisfies $K \geq K_{\mathrm{tr}}$, then the δ-approximate variant of Algorithm 1 with the trapezoid update rule* (4) *returns a 3ϵ-accurate regularization path.*

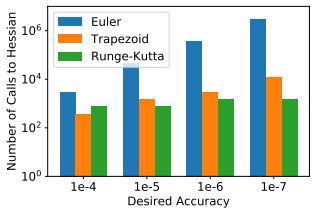
Building on Proposition 1, we propose second-order conjugate gradient (CG) versions of our methods that are inspired by Newton-CG methods (see, for example, [14] and [19] for a recent reference). Namely, we apply the CG method to compute the δ-accurate approximations of $v(\cdot,\cdot)$, which importantly only requires access to gradient and Hessian-vector product oracles. Corollary 1 summarizes the complexity of the CG variants of our methods in terms of the required number of gradient and Hessian-vector product oracle computations. The $\tilde{O}(\cdot)$ notation below hides logarithmic factors.

Corollary 1. *Under the same conditions as Proposition 1, suppose that we use the conjugate gradient method to compute the δ-accurate approximations of $v(\cdot,\cdot)$ where $\delta \in (0,\epsilon)$. Then, the Euler update variant of Algorithm 1 requires at most $\tilde{O}\left(\frac{((f(x_0)-f^*)L^{3/2}T}{\epsilon\bar{\mu}^{3/2}}\right)$ gradient and Hessian-vector product oracle calls to compute a 5ϵ-accurate regularization path, and the trapezoid variant of Algorithm 1 requires at most $\tilde{O}\left(\frac{L(1+G)^{3/2}T}{\sqrt{\epsilon\bar{\mu}}}\right)$ gradient and Hessian-vector product oracle calls to compute a 3ϵ-accurate regularization path.*

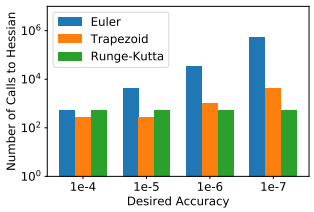
4 Computational experiments and results

Herein we examine the empirical behavior of each of the previously presented methods on logistic regression (LR) problems using the breast cancer dataset [3] (32 features and 569 observations). In particular, let $\{(a_i,b_i)\}_{i=1}^n$ denote a training set of features $a_i \in \mathbb{R}^p$ and labels $b_i \in \{-1,+1\}$ and define the sets of positive and negative examples by $S_+ := \{i \in [n] : b_i = 1\}$ and $S_- := \{i \in [n] : b_i = -1\}$. We examine two LR variants: *(i)* regularized LR with $f(x) = \frac{1}{n}\sum_{i=1}^n \log(1 + e^{-b_i a_i^T x})$, $\Omega(x) = \frac{1}{2}\|x\|^2$, $\lambda_{\min} = 10^{-4}$, $\lambda_{\max} = 10^4$, and *(ii)* re-weighted LR with $f(x) = \frac{1}{|S_+|}\sum_{i\in S_+} \log(1 + e^{-b_i a_i^T x})$, $\Omega(x) = \frac{1}{|S_-|}\sum_{i\in S_-} \log(1 + e^{-b_i a_i^T x})$, $\lambda_{\min} = 10^{-3}$, $\lambda_{\max} = 1$.

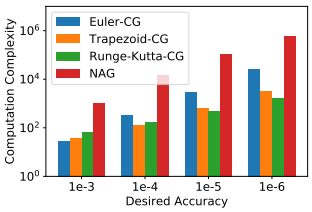
In parts *(a)* and *(b)* of Figure 1, we vary the desired accuracy parameter ϵ and plot the number of Hessian oracle computations required by the Euler, trapezoid, and Runge-Kutta variants of Algorithm 1. The total number of iterations K of each method is set according to a "doubling trick," whereby for each value of K we calcuate the observed accuracy along the path via interpolation and if the observed accuracy is too large then we double the value of K until it is below ϵ. Part *(c)* of Figure 1 performs a similar experiment for the conjugate gradient variants of the methods. In this case, we also compare with applying Nesterov's accelerated gradient method [11] to (1) independently at each of the grid of $\{\lambda_k\}$ values defined in the Euler update (3). Part *(c)* shows the total number of gradient and Hessian-vector product oracles for each of the four methods as we vary the desired accuracy parameter ϵ. All of the figures demonstrate the viability of our methods, as well as the superior performance of the trapezoid and Runge-Kutta methods due to the higher-order smoothness of the logistic loss. Part *(c)* also demonstrates the benefit of our differential equation approach over naively solving a series of optimization problems on a grid.



(a) Exact versions of Algorithm 1 applied to regularized LR.



(b) Exact versions of Algorithm 1 applied to re-weighted LR.



(c) CG versions of Algorithm 1 applied to re-weighted LR.

Figure 1

Acknowledgments

This research is supported by NSF Awards CCF-1755705 and CMMI-1762744.

References

[1] J. C. Butcher. *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods*. Wiley-Interscience, New York, NY, USA, 1987.

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[3] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[4] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[5] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[6] Walter Gautschi. *Numerical analysis*. Springer Science & Business Media, 2011.

[7] Joachim Giesen, Martin Jaggi, and Sören Laue. Approximating parameterized convex optimization problems. *ACM Transactions on Algorithms (TALG)*, 9(1):10, 2012.

[8] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct):1391–1415, 2004.

[9] David G Luenberger, Yinyu Ye, et al. Linear and nonlinear programming. *International Series in Operations Research and Management Science ReDIF-Book*, 2016.

[10] Eugene Ndiaye, Tam Le, Olivier Fercoq, Joseph Salmon, and Ichiro Takeuchi. Safe grid search with optimal complexity. In *International Conference on Machine Learning*, pages 4771–4780, 2019.

[11] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[12] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2018.

[13] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[14] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[15] Michael R Osborne, Brett Presnell, and Berwin A Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

[16] James Renegar. *A mathematical view of interior-point methods in convex optimization*, volume 3. Siam, 2001.

[17] Saharon Rosset. Following curved regularized optimization solution paths. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1153–1160. MIT Press, 2005.

[18] Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030, 2007.

[19] Clément W Royer, Michael O'Neill, and Stephen J Wright. A newton-cg algorithm with complexity guarantees for smooth unconstrained optimization. *Mathematical Programming*, pages 1–38, 2019.

[20] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[21] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

A Appendix

A.1 Proof of Theorem 1

The following lemma provides the computation guarantee of Taylor expansion.

Lemma 1 (Lemma 1 in [13])**.** *For a twice differentiable function $\phi(\cdot)$ with L-Lipschitz Hessian, it holds that*

$$\left\|\nabla\phi(y)-\nabla\phi(x)-\nabla^2\phi(x)(y-x)\right\|\le\frac{1}{2}L\left\|y-x\right\|^2,$$
$$\left\|\phi(y)-\phi(x)-\nabla\phi(x)^T(y-x)-\frac{1}{2}(y-x)^T\nabla^2\phi(x)(y-x)\right\|\le\frac{1}{6}L\left\|y-x\right\|^3.$$

Then we provide the local analysis of accuracy, namely $r_k:=\|\nabla f(x_k)+\lambda_k\nabla\Omega(x_k)\|$.

Lemma 2. *Under Assumption 1, discretization scheme* (3)*, it holds that*

$$r_{k+1}\le\left(1+h\cdot\frac{c_k}{\lambda_k}\right)r_k+h^2\cdot\frac{L_{f,2}+\lambda_{k+1}L_{\Omega,2}}{2}\left\|d_k\right\|^2.$$

Or equivalently

$$\frac{r_{k+1}}{\lambda_{k+1}}\le\frac{r_k}{\lambda_k}+h^2\cdot\frac{L_{f,2}+\lambda_{k+1}L_{\Omega,2}}{2\lambda_{k+1}}\left\|d_k\right\|^2. \tag{5}$$

Proof. From Lemma 1 and Assumption 1, it holds that

$$\left\|\nabla f(x_{k+1})-\nabla f(x_k)-\nabla^2 f(x_k)(x_{k+1}-x_k)\right\|\le\frac{L_{f,2}}{2}\left\|x_{k+1}-x_k\right\|^2 \tag{6}$$
$$\left\|\nabla\Omega(x_{k+1})-\nabla\Omega(x_k)-\nabla^2\Omega(x_k)(x_{k+1}-x_k)\right\|\le\frac{L_{\Omega,2}}{2}\left\|x_{k+1}-x_k\right\|^2 \tag{7}$$

Let $d_k=\left(\lambda_{k+1}\nabla^2\Omega(x_k)+\nabla^2 f(x_k)\right)^{-1}\frac{c_k}{\lambda_k}\nabla f(x_k)$. Also it holds that

$$\begin{aligned}
&\lambda_{k+1}\left(\nabla\Omega(x_k)+\nabla^2\Omega(x_k)(x_{k+1}-x_k)\right)+\nabla f(x_k)+\nabla^2 f(x_k)(x_{k+1}-x_k)\\
=&(\lambda_k\nabla\Omega(x_k)+\nabla f(x_k))+h\left(c_k\nabla\Omega(x_k)+\left(\lambda_{k+1}\nabla^2\Omega(x_k)+\nabla^2 f(x_k)\right)d_k\right)\\
=&(\lambda_k\nabla\Omega(x_k)+\nabla f(x_k))+h\left(c_k\nabla\Omega(x_k)+\frac{c_k}{\lambda_k}\nabla f(x_k)\right)\\
=&\left(1+h\cdot\frac{c_k}{\lambda_k}\right)(\lambda_k\nabla\Omega(x_k)+\nabla f(x_k)),
\end{aligned} \tag{8}$$

where the second equality holds since

$$d_k=\left(\lambda_{k+1}\nabla^2\Omega(x_k)+\nabla^2 f(x_k)\right)^{-1}\frac{c_k}{\lambda_k}\nabla f(x_k).$$

Combine (6), (7), (8), it holds that

$$\begin{aligned}
r_{k+1}&=\|\nabla f(x_{k+1})+\lambda_{k+1}\nabla\Omega(x_{k+1})\|\\
&\le\left\|\nabla f(x_{k+1})-\nabla f(x_k)-\nabla^2 f(x_k)(x_{k+1}-x_k)\right\|\\
&\quad+\left\|\lambda_{k+1}\left(\nabla\Omega(x_{k+1})-\nabla\Omega(x_k)-\nabla^2\Omega(x_k)(x_{k+1}-x_k)\right)\right\|\\
&\quad+\left\|\lambda_{k+1}\left(\nabla\Omega(x_k)+\nabla^2\Omega(x_k)(x_{k+1}-x_k)\right)+\nabla f(x_k)+\nabla^2 f(x_k)(x_{k+1}-x_k)\right\|\\
&\le\frac{L_{f,2}}{2}\left\|x_{k+1}-x_k\right\|^2+\frac{\lambda_{k+1}L_{\Omega,2}}{2}\left\|x_{k+1}-x_k\right\|^2\\
&\quad+\left(1+h\cdot\frac{c_k}{\lambda_k}\right)\|\nabla f(x_k)+\lambda_k\nabla\Omega(x_k)\|\\
&\le\left(1+h\cdot\frac{c_k}{\lambda_k}\right)r_k+h^2\cdot\frac{L_{f,2}+\lambda_{k+1}L_{\Omega,2}}{2}\left\|d_k\right\|^2.
\end{aligned} \tag{9}$$

□

The following lemma state a fact that under discretization scheme (3), the function value $f(x_k)$ will not increase. It implies that all near optimal solution x_k lie in the level set $S_{x_0} := \{x : f(x) \leq f(x_0)\}$.

Lemma 3. *Suppose $c_k < 0$, then under Assumption 1, discretization* (3)*, for all iterate $x_k \in S_{x_0}$ and step size h satisfying*

$$h \leq \min\left\{ \frac{\lambda_k}{-2c_k}, \sqrt{\frac{\frac{3\lambda_k}{-c_k}(\mu + \lambda_{k+1})}{L_{f,2}\left\|d_k\right\|}} \right\}, \tag{10}$$

where

$$G := \max_{x \in S_{x_0}} \left\|\nabla f(x)\right\|, \tag{11}$$

it holds that

$$f(x_{k+1}) \leq f(x_k) + \frac{h}{4} \cdot \frac{\lambda_k}{c_k} \cdot (\mu + \lambda_{k+1}) \left\|d_k\right\|^2 \leq f(x_k). \tag{12}$$

Proof. From Lemma 1, the function value at x_{k+1} can be upper bounded by

$$f(x_{k+1}) \leq f(x_k) + h \cdot \nabla f(x_k)^T d_k + h^2 \cdot \frac{1}{2} d_k^T \nabla^2 f(x_k) d_k + h^3 \cdot \frac{1}{6} L_{f,2} \left\|d_k\right\|^3. \tag{13}$$

For simplicity, we use the notation

$$H := \nabla^2 f(x_k), \quad \tilde{H} := H + \lambda_{k+1} \nabla^2 \Omega(x_k).$$

Apply discretization (3) to the inequality (13) we have

$$f(x_{k+1}) \leq f(x_k) + h \cdot \frac{\lambda_k}{c_k} \cdot d_k^T \tilde{H} d_k + h^2 \cdot \frac{1}{2} d_k^T H d_k + h^3 \cdot \frac{1}{6} L_{f,2} \left\|d_k\right\|^3. \tag{14}$$

Then we only need an upper bound of high order terms with respect to h. Since the step size h satisfies the condition (10), it holds that $h + \frac{\lambda_k}{2c_k} \leq 0$. Combine with the property that $H \preceq \tilde{H}$ and $c_k < 0$, we have

$$\frac{h}{4} \cdot \frac{\lambda_k}{c_k} \cdot d_k^T \tilde{H} d_k + h^2 \cdot \frac{1}{2} d_k^T H d_k \leq 0.$$

Also, (10) leads to

$$h^2 \leq \frac{\frac{3\lambda_k}{-c_k}(\mu + \lambda_{k+1})}{L_{f,2}\left\|d_k\right\|}.$$

Combine with the fact that $\tilde{H} \succeq (\mu + \lambda_{k+1}) I$ and $c_k < 0$ it holds that

$$\frac{h}{2} \cdot \frac{\lambda_k}{c_k} \cdot d_k^T \tilde{H} d_k + h^3 \cdot \frac{1}{6} L_{f,2} \left\|d_k\right\|^3 \leq \frac{h}{2} \cdot \frac{\lambda_k}{c_k} \cdot (\mu + \lambda_{k+1}) \left\|d_k\right\|^2 - \frac{h}{2} \cdot \frac{\lambda_k}{c_k} \cdot (\mu + \lambda_{k+1}) \left\|d_k\right\|^2 = 0.$$

Combine the inequalities above, we hence have

$$f(x_{k+1}) \leq f(x_k) + \frac{h}{4} \cdot \frac{\lambda_k}{c_k} \cdot d_k^T \tilde{H} d_k \leq f(x_k).$$

□

As a by-product of Lemma 3, we provide an upper bound on the norm of d_k.

Proposition 2. *Suppose h satisfies* (10) *and $c_k < 0$, then it holds that*

$$h \left\|d_k\right\|^2 \leq \frac{4(f(x_k) - f(x_{k+1}))}{\frac{\lambda_k}{-c_k} \cdot (\mu + \lambda_{k+1})}. \tag{15}$$

Proof. Since h satisfies (10), by Lemma 3 it holds that

$$f(x_{k+1}) \leq f(x_k) + \frac{h}{4} \cdot \frac{\lambda_k}{c_k} \cdot (\mu + \lambda_{k+1}) \left\|d_k\right\|^2. \tag{16}$$

With $c_k < 0$, (16) is equivalent to

$$\left\|d_k\right\|^2 \leq \frac{4(f(x_k) - f(x_{k+1}))}{\frac{\lambda_k}{-c_k} \cdot (\mu + \lambda_{k+1})}.$$

□

The follow lemma provides the global analysis, that is a uniform upper bound on all r_k.

Lemma 4. *Suppose $\lambda_{k+1} \geq \lambda_{\min}$, $c_j < 0$ for all $j \leq k$, step-size h satisfies*

$$h \leq \min\left\{\frac{\lambda_j}{-2c_j}, \sqrt{\frac{\frac{3\lambda_j}{-c_j}(\mu+\lambda_{j+1})^2}{L_{f,2}G}}\right\}, \forall j \leq k. \tag{17}$$

Then under Assumption 1, discretization (3)*, it holds that*

$$\frac{r_{k+1}}{\lambda_{k+1}} \leq \frac{r_0}{\lambda_0} + 2h\left(\frac{L_{f,2}}{\Lambda_{1,k}} + \frac{L_{\Omega,2}}{\Lambda_{2,k}}\right)(f(x_0) - f(x_{k+1})), \tag{18}$$

where

$$\Lambda_{1,k} := \min_{j\leq k}\left\{\frac{\lambda_{j+1}\lambda_j}{-c_j}\cdot(\mu+\lambda_{j+1})\right\}, \quad \Lambda_{2,k} := \min_{j\leq k}\left\{\frac{\lambda_j}{-c_j}\cdot(\mu+\lambda_{j+1})\right\}.$$

Proof. First suppose $x_j \in S_{x_0}$ and it leads to $\|d_j\| \leq \frac{G}{\mu+\lambda_{j+1}}$. Since h satisfies (17) it holds that $f(x_{j+1}) \leq f(x_j)$ and it implies that $x_{j+1} \in S_{x_0}$. Then by induction we conclude that $x_j \in S_{x_0}$ for all $j \leq k$. Hence (17) implies (10). Then by Proposition 2 for all $j \leq k$ it holds that

$$h\|d_j\|^2 \leq \frac{4(f(x_j) - f(x_{j+1}))}{\frac{\lambda_j}{-c_j}\cdot(\mu+\lambda_{j+1})}. \tag{19}$$

Apply (19) to Lemma 2, then for all $j \leq k$ it holds that

$$\begin{aligned}\frac{r_{j+1}}{\lambda_{j+1}} &\leq \frac{r_j}{\lambda_j} + h^2\cdot\frac{L_{f,2}+\lambda_{j+1}L_{\Omega,2}}{2\lambda_{j+1}}\|d_j\|^2 \\ &\leq \frac{r_j}{\lambda_j} + h\cdot\frac{L_{f,2}+\lambda_{j+1}L_{\Omega,2}}{2\lambda_{j+1}}\cdot\frac{4(f(x_j)-f(x_{j+1}))}{\frac{\lambda_j}{-c_j}\cdot(\mu+\lambda_{j+1})}\end{aligned}$$

Taking the summation over j from 0 to k, we get

$$\begin{aligned}\frac{r_{k+1}}{\lambda_{k+1}} &\leq \frac{r_0}{\lambda_0} + h\sum_{j=0}^{k}\frac{L_{f,2}+\lambda_{j+1}L_{\Omega,2}}{2\lambda_{j+1}}\cdot\frac{4(f(x_j)-f(x_{j+1}))}{\frac{\lambda_j}{-c_j}\cdot(\mu+\lambda_{j+1})} \\ &= \frac{r_0}{\lambda_0} + 2h\sum_{j=0}^{k}L_{f,2}\cdot\frac{(f(x_j)-f(x_{j+1}))}{\frac{\lambda_{j+1}\lambda_j}{-c_j}\cdot(\mu+\lambda_{j+1})} + L_{\Omega,2}\cdot\frac{(f(x_j)-f(x_{j+1}))}{\frac{\lambda_j}{-c_j}\cdot(\mu+\lambda_{j+1})}.\end{aligned}$$

By definition of $\Lambda_{1,k}, \Lambda_{2,k}$, we can further get

$$\begin{aligned}\frac{r_{k+1}}{\lambda_{k+1}} &\leq \frac{r_0}{\lambda_0} + 2h\sum_{j=0}^{k}\left(\frac{L_{f,2}}{\Lambda_{1,k}} + \frac{L_{\Omega,2}}{\Lambda_{2,k}}\right)(f(x_j) - f(x_{j+1})) \\ &= \frac{r_0}{\lambda_0} + 2h\left(\frac{L_{f,2}}{\Lambda_{1,k}} + \frac{L_{\Omega,2}}{\Lambda_{2,k}}\right)(f(x_0) - f(x_{k+1})).\end{aligned}$$

□

The following theorem considers the case when $-\lambda \leq \xi(\lambda) < 0$, which is a more general scenario than the scenario of interest $\xi(\lambda) = -\lambda$.

Theorem 3. *Suppose $\lambda_{k+1} \geq \lambda_{\min}$, $-\lambda \leq \xi(\lambda) < 0$ for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and step-size h satisfies*

$$h \leq \min\left\{\frac{1}{2}, \sqrt{\frac{3(\mu+\lambda_{\min})^2}{L_{f,2}G}}\right\}. \tag{20}$$

Let f^* denote the minimum value of $f(\cdot)$. Then under Assumption 1, discretization (3)*, it holds that*

$$r_{k+1} \leq r_0 + 2h\left(\frac{L_{f,2}}{\mu+\lambda_{\min}} + L_{\Omega,2}\right)(f(x_0) - f(x_{k+1})). \tag{21}$$

Proof. Since $c_j = \xi(\lambda_j) \in [-\lambda_j, 0)$, we have $\frac{\lambda_j}{-c_j} \geq 1$. Therefore it holds that

$$\frac{\lambda_j}{-2c_j} \geq \frac{1}{2} \quad \text{and} \quad \sqrt{\frac{\frac{3\lambda_j}{-c_j}(\mu+\lambda_{j+1})^2}{L_{f,2}G}} \geq \sqrt{\frac{3(\mu+\lambda_{\min})^2}{L_{f,2}G}},$$

and hence condition (20) implies condition (17). Furthermore, we have

$$\Lambda_{1,k} := \min_{j\leq k}\left\{\frac{\lambda_{j+1}\lambda_j}{-c_j}\cdot(\mu+\lambda_{j+1})\right\} \geq \min_{j\leq k}\left\{\lambda_{j+1}(\mu+\lambda_{j+1})\right\} = \lambda_{k+1}(\mu+\lambda_{k+1}) \tag{22}$$

and

$$\Lambda_{2,k} := \min_{j\leq k}\left\{\frac{\lambda_j}{-c_j}\cdot(\mu+\lambda_{j+1})\right\} \geq \min_{j\leq k}\left\{(\mu+\lambda_{j+1})\right\} = \mu+\lambda_{k+1}. \tag{23}$$

Apply inequalities (22) and (23) to (18) we obtain

$$\begin{aligned} r_{k+1} &\leq r_0 + 2h\left(\frac{L_{f,2}}{\mu+\lambda_{k+1}} + L_{\Omega,2}\right)(f(x_0) - f(x_{k+1})) \\ &\leq r_0 + 2h\left(\frac{L_{f,2}}{\mu+\lambda_{\min}} + L_{\Omega,2}\right)(f(x_0) - f^*). \end{aligned}$$

□

Now we introduce the linear interpolation of $\{\lambda_k\}$ and $\{x_k\}$, which is formally defined as

$$\begin{cases} \hat{\lambda}(t) := \alpha\lambda_j + (1-\alpha)\lambda_{j+1} \\ \hat{x}(t) := \alpha x_j + (1-\alpha)x_{j+1}, \end{cases} \tag{24}$$

where

$$\alpha := \frac{t_{j+1}-t}{h}, t \in [t_j, t_{j+1}].$$

The following theorem provides the computation guarantee of linear interpolation.

Theorem 4. *When step-size h satisfies* (10)*, for all $t \in [t_0, t_k]$, it holds that*

$$\left\|\nabla F_{\hat{\lambda}(t)}(\hat{x}(t))\right\| \leq r_{\max} + h^2\cdot\left(\frac{L_{f,2}}{8}D_1 + \frac{L_{\Omega,2}}{8}D_2 + \frac{L_{\Omega,1}}{4}D_3\right),$$

where

$$r_{\max} := \max_{0\leq j\leq k} r_j,\ D_1 := \max_{0\leq j\leq k}\|d_j\|^2,\ D_2 := \max_{0\leq j\leq k}\lambda_j\|d_j\|^2,\ D_3 := \max_{0\leq j\leq k}|c_j|\,\|d_j\|.$$

Proof. Suppose $t \in [t_j, t_{j+1}]$. For simplicity let $x := \hat{x}(t)$, $\lambda := \hat{\lambda}(t)$, $\delta_1 := \nabla f(x_j) + \lambda_j\nabla\Omega(x_j)$, $\delta_2 := \nabla f(x_{j+1}) + \lambda_{j+1}\nabla\Omega(x_{j+1})$. First we have

$$\begin{aligned} \left\|\nabla f(x_j) - \nabla f(x) - \nabla^2 f(x)\cdot(x_j - x)\right\| &\leq \frac{L_{f,2}}{2}\|x_j - x\|^2, \\ \left\|\nabla f(x_{j+1}) - \nabla f(x) - \nabla^2 f(x)\cdot(x_{j+1} - x)\right\| &\leq \frac{L_{f,2}}{2}\|x_{j+1} - x\|^2. \end{aligned}$$

Combine the above two inequality we have

$$\begin{aligned} &\|\alpha\nabla f(x_j) + (1-\alpha)\nabla f(x_{j+1}) - \nabla f(x)\| \\ \leq &\frac{L_{f,2}}{2}\left(\alpha\|x_j - x\|^2 + (1-\alpha)\|x_{j+1} - x\|^2\right) \\ = &\frac{\alpha(1-\alpha)L_{f,2}}{2}\|x_j - x_{j+1}\|^2 \leq \frac{L_{f,2}h^2}{8}\|d_j\|^2. \end{aligned}$$

Then,

$$\begin{aligned}
&\|\lambda\nabla\Omega(x) - \alpha\lambda_j\nabla\Omega(x_j) - (1-\alpha)\lambda_{j+1}\nabla\Omega(x_{j+1})\| \\
\leq &\|\lambda(\alpha\nabla\Omega(x_j) + (1-\alpha)\nabla\Omega(x_{j+1}) - \nabla\Omega(x))\| \\
&+ \|\alpha(\lambda-\lambda_j)\nabla\Omega(x_j) + (1-\alpha)(\lambda-\lambda_{j+1})\nabla\Omega(x_{j+1})\| \\
\leq &\frac{\lambda L_{\Omega,2}}{2}\left(\alpha\|x_j - x\|^2 + (1-\alpha)\|x_{j+1}-x\|^2\right) \\
&+ \|\alpha(1-\alpha)(\lambda_{j+1}-\lambda_j)(\nabla\Omega(x_{j+1}) - \nabla\Omega(x_j))\| \\
\leq &\frac{\lambda L_{\Omega,2}}{2}\alpha(1-\alpha)\|x_{j+1}-x_j\|^2 + \|\alpha(1-\alpha)(\lambda_{j+1}-\lambda_j)(\nabla\Omega(x_{j+1}) - \nabla\Omega(x_j))\| \\
\leq &\frac{\lambda L_{\Omega,2}}{8}\|h\cdot d_j\|^2 + \frac{h|c_j|L_{\Omega,1}}{4}\|x_{j+1}-x_j\| \\
= &h^2\cdot\left(\frac{\lambda L_{\Omega,2}}{8}\|d_j\|^2 + \frac{|c_j|L_{\Omega,1}}{4}\|d_j\|\right).
\end{aligned}$$

Combine the above two inequality it holds that

$$\begin{aligned}
&\|\nabla f(x) + \lambda x - \alpha\delta_1 - (1-\alpha)\delta_2\| \\
\leq &\frac{L_{f,2}h^2}{8}\|d_j\|^2 + h^2\cdot\left(\frac{\lambda L_{\Omega,2}}{8}\|d_j\|^2 + \frac{|c_j|L_{\Omega,1}}{4}\|d_j\|\right) \\
\leq &h^2\cdot\left(\frac{L_{f,2}}{8}D_1 + \frac{L_{\Omega,2}}{8}D_2 + \frac{L_{\Omega,1}}{4}D_3\right).
\end{aligned}$$

From definition of $r_{\max}$ we know that

$$\max\{\|\delta_1\|, \|\delta_2\|\} \leq r_{\max}.$$

Then, it holds that

$$\|\nabla f(x) + \lambda x\| \leq r_{\max} + h^2\cdot\left(\frac{L_{f,2}}{8}D_1 + \frac{L_{\Omega,2}}{8}D_2 + \frac{L_{\Omega,1}}{4}D_3\right).$$

□

Then we come to the scenario $\xi(\lambda) = -\lambda$. The following proposition provides the computation guarantee of path accuracy in terms of step-size h.

Proposition 3. *Let R_f denote $f(x_0) - F^*$. When step-size h satisfies* (20)*, under Assumption 1, discretization* (3)*, for all $t \in [t_0, t_k]$, it holds that*

$$\left\|\nabla F_{\hat{\lambda}(t)}(\hat{x}(t))\right\| \leq r_0 + 2h\left(\frac{L_{f,2}}{\tilde{\mu}} + L_{\Omega,2}\right)R_f + h^2\cdot\left(\frac{L_{f,2}G^2}{8\tilde{\mu}^2} + \frac{L_{\Omega,2}G^2}{8\tilde{\mu}} + \frac{L_{\Omega,1}G}{4}\right).$$

Proof. We have condition (20) implies condition (10). Then apply results in Theorem 4, it holds that

$$\left\|\nabla F_{\hat{\lambda}(t)}(\hat{x}(t))\right\| \leq r_{\max} + h^2\cdot\left(\frac{L_{f,2}}{8}D_1 + \frac{L_{\Omega,2}}{8}D_2 + \frac{L_{\Omega,1}}{4}D_3\right),$$

where

$$r_{\max} = \max_{0\leq j\leq k} r_j \leq r_0 + 2h\left(\frac{L_{f,2}}{\tilde{\mu}} + L_{\Omega,2}\right)R_f$$

and

$$D_1 = \max_{0\leq j\leq k}\|d_j\|^2 \leq \frac{G^2}{\tilde{\mu}^2}, \quad D_2 = \max_{0\leq j\leq k}\lambda_j\|d_j\|^2 \leq \frac{G^2}{\tilde{\mu}}, \quad D_3 = \max_{0\leq j\leq k}|c_j|\,\|d_j\| \leq G.$$

□

Finally we provide a theorem on oracle complexity of Algorithm 1 with update scheme (3) that and also the proof of Theorem 1.

Theorem 5. *Given a μ-strongly convex function $f(\cdot)$ and a 1-strongly convex function $\Omega(\cdot)$ satisfying Assumption 1, constant $\lambda_{\max} > \lambda_{\min} > 0$, $\epsilon > 0$, an initialization x_0. Let K denote number of iterations such that $K \geq \max\left\{2T, \frac{\sqrt{LG}T}{\sqrt{3\tilde{\mu}}}\right\}$. Algorithm 1 with update scheme* (3) *guarantees that*

$$\max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \|\nabla F_\lambda(\hat{x}(\lambda))\| \leq r_0 + \frac{2T}{K}\cdot R_f L\left(\frac{1}{\tilde{\mu}}+1\right) + \frac{T^2}{4K^2}\cdot L(G+1)^2\left(\frac{1}{\tilde{\mu}}+1\right)^2, \quad (25)$$

where r_0 is the initial error $r_0 := \|\nabla F_{\lambda_{\max}}(x_0)\|$, $T = \ln\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)$, $L = \max\{L_{f,2}, L_{\Omega,1}, L_{\Omega,2}\}$.

Proof of Theorem 5. Since $\lambda_0 = \lambda_{\max}$, $\lambda_K = \lambda_{\min}$ and $\lambda_K = \lambda_0\cdot(1-h)^K$, it holds that

$$h = 1 - \left(\frac{\lambda_{\min}}{\lambda_{\max}}\right)^{1/K} \leq \frac{T}{K} \leq \frac{T}{K_0} \leq \left\{\frac{1}{2}, \frac{\sqrt{3\tilde{\mu}}}{\sqrt{L_{f,2}G}}\right\}$$

and hence condition (20) is satisfied. Then by Proposition 3, it holds that

$$\begin{aligned}
\left\|\nabla F_{\hat{\lambda}(t)}(\hat{x}(t))\right\| &\leq r_0 + 2h\left(\frac{L_{f,2}}{\tilde{\mu}} + L_{\Omega,2}\right)R_f + h^2\cdot\left(\frac{L_{f,2}G^2}{8\tilde{\mu}^2} + \frac{L_{\Omega,2}G^2}{8\tilde{\mu}} + \frac{L_{\Omega,1}G}{4}\right)\\
&\leq r_0 + \frac{2T}{K}\cdot\left(\frac{L}{\tilde{\mu}}+L\right)R_f + \left(\frac{T}{K}\right)^2\cdot\left(\frac{LG^2}{8\tilde{\mu}^2} + \frac{LG^2}{8\tilde{\mu}} + \frac{LG}{4}\right)\\
&\leq r_0 + \frac{2T}{K}\cdot R_f L\left(\frac{1}{\tilde{\mu}}+1\right) + \frac{T^2}{4K^2}\cdot L(G+1)^2\left(\frac{1}{\tilde{\mu}}+1\right)^2.
\end{aligned}$$

□

Proof of Theorem 1. The condition that $K \geq \max\left\{2T, \frac{\sqrt{LG}T}{\sqrt{3\tilde{\mu}}}\right\}$ guarantees that step-size h satisfies (20). Also $K \geq \frac{R_f L(1+\tilde{\mu})T}{\epsilon\tilde{\mu}}$ and $K \geq \frac{\sqrt{L}(G+1)(1+\tilde{\mu})T}{\sqrt{\epsilon\tilde{\mu}}}$ guarantees that $\frac{2T}{K}\cdot R_f L\left(\frac{1}{\tilde{\mu}}+1\right) \leq 2\epsilon$ and $\frac{T^2}{4K^2}\cdot L(G+1)^2\left(\frac{1}{\tilde{\mu}}+1\right)^2 \leq \epsilon$. Hence Algorithm 1 with (3) guarantees a 4ϵ-accurate regularization path. □

A.2 Proof of Theorem 2

First we introduce the directional derivative. For $p \geq 1$, let $D^p f(x)[h_1, \ldots, h_p]$ denote the directional derivative of function f at x along directions $h_i, i = 1, \ldots, p$. For instance, $Df(x)[h] = \nabla f(x)^T h$ and $D^2 f(x)[h_1, h_2] = h_1^T \nabla^2 f(x) h_2$. Also, the norm of directional derivatives is defined as

$$\|D^p f(x)\| := \max_{h_1,\ldots,h_p}\left\{|D^p f(x)[h_1,\ldots,h_p]| : \|h_i\| \leq 1\right\}.$$

For detailed properties of directional derivatives we refer readers to [12]. Then we state the main assumption and notation.

Assumption 2. *There exist constants $L_{f,i}$ and $L_{\Omega,i}$ such that for all $x, y \in \mathbb{R}^n$ and $i \in \{1,2,3\}$, it holds that*

$$\begin{aligned}
\left\|D^i f(x) - D^i f(y)\right\| &\leq L_{f,i}\|x-y\|,\\
\left\|D^i \Omega(x) - D^i \Omega(y)\right\| &\leq L_{\Omega,i}\|x-y\|.
\end{aligned}$$

Definition 2. $G = \max\{G_f, G_\Omega\}$. $L := \max\{L_{f,i}, L_{\Omega,i}\}$. $\tilde{\mu}_x := \sigma_{\min}\left(\nabla^2 f(x)\right) + (1-h)\lambda\cdot\sigma_{\min}\left(\nabla^2\Omega(x)\right)$.

As a direct consequence of the Assumption 2, the following lemma guarantees the accuracy of Taylor expansion.

Lemma 5. *For all function f with an L_n-Lipschitz n-th derivative and $x, y \in \mathbb{R}^p$ it holds that*

$$\left|f(y) - \sum_{i=0}^{n} D^i f(x)[y-x]^i\right| \leq \frac{L_n}{n!}\|x-y\|^n.$$

We make an technical assumption on the step-size that we will validate later.

Assumption 3. *Suppose step-size h satisfies that*

$$h \leq \min\left\{0.2, \frac{\tilde{\mu}_{\min}}{8L(1+G)}\right\}.$$

The following two technical lemma provide an upper bound on $\|d_k\|$ and $\|d_{k,1} - d_{k,2}\|$.

Lemma 6. *Suppose $x \in S_{x_0}, \lambda \in [\lambda_{\min}, \lambda_{\max}], h > 0$ and $(\tilde{x}, \tilde{\lambda}) = T(x, \lambda; h)$. Let r denote the initial residual $\|\nabla f(x) + \lambda \nabla \Omega(x)\|$ satisfying that $r \leq \tilde{\mu}$. Then it holds that*

$$\|d_1\| \leq \frac{2(G+1)}{1+\lambda}. \tag{26}$$

Proof. Let $H = \nabla^2 f(x) + \lambda \nabla^2 \Omega(x)$. By definition $d_1 = -H^{-1}\nabla f(x)$. When $\lambda \geq 1$, it holds that $(1+\lambda)\|d_1\| \leq \frac{1+\lambda}{\lambda} G_f \leq 2G_f$. Also when $\lambda \leq 1$, it holds that

$$\begin{aligned}
&(1+\lambda)\|d_1\| \leq (1+\lambda)\left(\left\|H^{-1}\lambda\nabla\Omega(x)\right\| + \left\|H^{-1}(\nabla f(x) + \lambda\nabla\Omega(x))\right\|\right) \\
\leq& 2\left(G_\Omega + \frac{r}{\tilde{\mu}}\right) \leq 2(G_\Omega + 1).
\end{aligned}$$

□

Lemma 7. *Let $x \in \mathbb{R}^n$, $\lambda > 0$, x_i, λ_i, d_i generated by* (4)*. Under Assumption 2 and 3, it holds that*

$$\left\|\tilde{H}_1(d_2 - d_1) - \nabla^2 f(x_1)(x_1 - x_2) - (\tilde{H}_1 - \tilde{H}_2)d_2\right\| \leq \frac{L_{f,2}}{2}\|x_1 - x_2\|^2,$$

where $\tilde{H}_i = \nabla^2 f(x_i) + \lambda_i \nabla^2 \Omega(x_i), i \in \{1, 2\}$. Furthermore, it holds that

$$\|d_2 - d_1\| \leq h \cdot \frac{3L(1+\lambda)\left(\|d_1\| + \|d_1\|^2\right)}{\tilde{\mu}_x},$$

Proof. Using the definition of d_1, d_2 we have

$$\begin{aligned}
\tilde{H}_1(d_2 - d_1) &= \nabla f(x_1) - \tilde{H}_1\tilde{H}_2^{-1}\nabla f(x_2) \\
&= \nabla f(x_1) - \nabla f(x_2) + \left(I - \tilde{H}_1\tilde{H}_2^{-1}\right)\nabla f(x_2) \\
&= \nabla^2 f(x_1)(x_1 - x_2) + (R) + (\tilde{H}_2 - \tilde{H}_1)\tilde{H}_2^{-1}\nabla f(x_2) \\
&= \nabla^2 f(x_1)(x_1 - x_2) + (\tilde{H}_1 - \tilde{H}_2)d_2 + (R),
\end{aligned} \tag{27}$$

where

$$\begin{aligned}
\|(R)\| &= \left\|\nabla f(x_1) - \nabla f(x_2) - \nabla^2 f(x_1)(x_1 - x_2)\right\| \\
&\leq \frac{L_{f,2}}{2}\|x_1 - x_2\|^2 = \frac{h^2}{2} \cdot L_{f,2}\|d_1\|^2.
\end{aligned}$$

Also, it holds that

$$\left\|\nabla^2 f(x_1)(x_1 - x_2)\right\| = h\left\|\nabla^2 f(x_1)d_1\right\| \leq hL_{f,1}\|d_1\|,$$

and that

$$\begin{aligned}
&\|(\tilde{H}_1 - \tilde{H}_2)d_2\| \\
=& \left\|\left(\nabla^2 f(x_1) - \nabla^2 f(x_2) + \lambda_1(\nabla^2\Omega(x_1) - \nabla^2\Omega(x_2)) + (\lambda_1 - \lambda_2)\nabla^2\Omega(x_2)\right)d_2\right\| \\
\leq& \left(L_{f,2}\|x_1 - x_2\| + \lambda_1 L_{\Omega,2}\|x_1 - x_2\| + |\lambda_1 - \lambda_2| L_{\Omega,1}\right)\|d_2\| \\
\leq& h\left(\left(L_{f,2} + \lambda L_{\Omega,2}\right)\|d_1\| + \lambda L_{\Omega,1}\right)\|d_2\|.
\end{aligned}$$

Hence, it holds that

$$\begin{aligned}
&\tilde{\mu}_x\|d_2\| - \tilde{\mu}_x\|d_1\| \leq \tilde{\mu}_x\|d_2 - d_1\| \leq \left\|\tilde{H}_1(d_2 - d_1)\right\| \\
\leq& \frac{h^2}{2} \cdot L_{f,2}\|d_1\|^2 + hL_{f,1}\|d_1\| + h\left(\left(L_{f,2} + \lambda L_{\Omega,2}\right)\|d_1\| + \lambda L_{\Omega,1}\right)\|d_2\|,
\end{aligned} \tag{28}$$

where $\tilde{\mu}_x = \sigma_{\min}\left(\nabla^2 f(x_1)\right) + \lambda_1$. When h satisfies Assumption 3, it holds that

$$h\left(\left(L_{f,2} + \lambda L_{\Omega,2}\right)\|d_1\| + \lambda L_{\Omega,1}\right) \leq \frac{\tilde{\mu}_x}{3}, \quad \frac{h^2}{2} \cdot L_{f,2}\|d_1\| + hL_{f,1} \leq \frac{\tilde{\mu}_x}{3}. \tag{29}$$

Apply (29) to (28), it holds that $\frac{2}{3}\tilde{\mu}_1\|d_2\| \leq \frac{4}{3}\tilde{\mu}_1\|d_1\|$ and it implies that $\|d_2\| \leq 2\|d_1\|$. Apply this to (28), it holds that

$$\begin{aligned}\left\|\tilde{H}_1(d_1 - d_2)\right\| &\leq \frac{h^2}{2} \cdot L_{f,2}\|d_1\|^2 + hL_{f,1}\|d_1\| + h\left(\left(L_{f,2} + \lambda_0 L_{\Omega,2}\right)\|d_1\| + \lambda L_{\Omega,1}\right)\|d_2\| \\ &\leq 2hL(1+\lambda)\left(\|d_1\| + \|d_1\|^2\right).\end{aligned}$$

Hence, it holds that

$$\|d_1 - d_2\| \leq h \cdot \frac{2L(1+\lambda)\left(\|d_1\| + \|d_1\|^2\right)}{\tilde{\mu}_x}.$$

□

The following theorem provides the local accuracy analysis of update scheme 4.

Theorem 6. *Suppose $x \in S_{x_0}, \lambda \in [\lambda_{\min}, \lambda_{\max}], h > 0$ and $(\tilde{x}, \tilde{\lambda}) = T(x, \lambda; h)$. Let $r = \|\nabla f(x) + \lambda \nabla \Omega(x)\|$ satisfying that $r \leq \tilde{\mu}_x$. Then for all $x \in \mathbb{R}^p$, $\lambda > 0$ and $(\tilde{x}, \tilde{\lambda}) = T(x, \lambda)$, it holds that*

$$\tilde{r} := \left\|\nabla F_{\tilde{\lambda}}(\tilde{x})\right\| \leq \frac{\tilde{\lambda}}{\lambda}\|\nabla F_\lambda(x)\| + h^3 \cdot 3L(1+G)^3 + h^4 \cdot \frac{2L^3(1+G)^4}{\tilde{\mu}_x^2}. \tag{30}$$

Proof of Theorem 6. We will begin the local residual analysis by estimating the difference between $\nabla F_{\tilde{\lambda}}(\tilde{x})$ and $\frac{\tilde{\lambda}}{\lambda} \cdot \nabla F_\lambda(x)$. After rearrangement, we have

$$\begin{aligned}R &= \nabla F_{\tilde{\lambda}}(\tilde{x}) - \frac{\tilde{\lambda}}{\lambda} \cdot \nabla F_\lambda(x) \\ &= \nabla f(\tilde{x}) + \tilde{\lambda}\nabla\Omega(\tilde{x}) - \frac{\tilde{\lambda}}{\lambda}\left(\nabla f(x) + \lambda\nabla\Omega(x)\right) \\ &= \underbrace{\nabla f(\tilde{x}) - \nabla f(x)}_{(A)} + \underbrace{\tilde{\lambda}(\nabla\Omega(\tilde{x}) - \nabla\Omega(x))}_{(B)} + \underbrace{\left(1 - \frac{\tilde{\lambda}}{\lambda}\right)\nabla f(x)}_{(C)}.\end{aligned}$$

We will approach the result in (30) by splitting and rearranging terms in (A), (B) and (C). From Lemma 5, it holds that

$$\|(RA)\| := \|(A) - \underbrace{\nabla^2 f(x)(\tilde{x} - x)}_{(A')} - \underbrace{\frac{1}{2}D^3 f(x)[\tilde{x} - x]^2}_{(A3)}\| \leq \frac{L_{f,3}}{6}\|\tilde{x} - x\|^3 = \frac{h^3}{6}L_{f,3}\|\tilde{d}\|^3.$$

From the update (4), it holds that

$$(A') = \nabla^2 f(x_2)(\tilde{x} - x) = \nabla^2 f(x)\frac{h}{2}(d_1 + d_2) = \underbrace{h\nabla^2 f(x)d_1}_{(A1)} + \underbrace{\frac{h}{2}\nabla^2 f(x)(d_2 - d_1)}_{(A2)}.$$

Also, for (B) using Lemma 5 and based on update (4) we have

$$\begin{aligned}(B) &= \tilde{\lambda}(\nabla\Omega(\tilde{x}) - \nabla\Omega(x)) \\ &= \tilde{\lambda}\nabla^2\Omega(x)(\tilde{x}-x) + \underbrace{\tilde{\lambda}\cdot\frac{1}{2}D^3\Omega(x)[\tilde{x}-x]^2}_{(B4)} + (RB) \\ &= \lambda\nabla^2\Omega(x)(\tilde{x}-x) + \underbrace{(\tilde{\lambda}-\lambda)\nabla^2\Omega(x)(\tilde{x}-x)}_{(B3)} + (B4) + (RB) \\ &= \lambda\nabla^2\Omega(x)\frac{h}{2}(d_1+d_2) + (B3) + (B4) + (RB) \\ &= \underbrace{h\lambda\nabla^2\Omega(x)d_1}_{(B1)} + \underbrace{\frac{h}{2}\lambda\nabla^2\Omega(x)(d_2-d_1)}_{(B2)} + (B3) + (B4) + (RB),\end{aligned}$$

where

$$\begin{aligned}\|(RB)\| &= \tilde{\lambda}\left\|\nabla\Omega(\tilde{x}) - \nabla\Omega(x) - \nabla^2\Omega(x)(\tilde{x}-x) - \frac{1}{2}D^3\Omega(x)[\tilde{x}-x]^2\right\| \\ &\leq \frac{\tilde{\lambda}L_{\Omega,3}}{6}\|\tilde{x}-x\|^3 = \frac{h^3\tilde{\lambda}L_{\Omega,3}}{6}\left\|\tilde{d}\right\|^3.\end{aligned}$$

And for (C) we have

$$(C) = \left(h - \frac{h^2}{2}\right)\nabla f(x) = \underbrace{h\nabla f(x)}_{(C1)} - \underbrace{\frac{h^2}{2}\nabla f(x)}_{(C2)}.$$

Then we rearrange these terms to achieve $O(h^3)$ local residual as follows. First by the definition of d_1, we have

$$\begin{aligned}(A1)+(B1)+(C1) &= h\nabla^2 f(x)d_1 + \left(h-h^2\right)\lambda\nabla^2\Omega(x)d_1 + h\nabla f(x) \\ &= h\left(\nabla^2 f(x) + (1-h)\lambda\nabla^2\Omega(x)\right)d_1 + h\nabla f(x) \\ &= -h\nabla f(x) + h\nabla f(x) = 0.\end{aligned} \tag{31}$$

Using Lemma 7, we have

$$\begin{aligned}&(A2)+(B2) = \frac{h}{2}\nabla^2 f(x)(d_2-d_1) + \frac{h}{2}\lambda\nabla^2\Omega(x)(d_2-d_1) \\ =&\frac{h}{2}\left(\nabla^2 f(x) + \lambda\nabla^2\Omega(x)\right)(d_2-d_1) = \frac{h}{2}\tilde{H}_0(d_2-d_1) \\ =&\underbrace{\frac{h}{2}\nabla^2 f(x)(x_1-x_2)}_{(D1)} + \underbrace{\frac{h}{2}(\nabla^2 f(x_1)-\nabla^2 f(x_2))d_2}_{(D2)} + \underbrace{\frac{h}{2}(\lambda_1\nabla^2\Omega(x_1)-\lambda_2\nabla^2\Omega(x_2))d_2}_{(D3)} + (RD),\end{aligned} \tag{32}$$

where

$$\|(RD)\| \leq \frac{h}{2}\cdot\frac{L_{f,2}}{2}\|x_1-x_2\|^2 = \frac{h^3}{4}L_{f,2}\|d_1\|^2.$$

Furthermore, it holds that

$$(D1)+(C2) = -\frac{h^2}{2}\nabla^2 f(x)d_1 - \frac{h^2}{2}\nabla f(x) = \underbrace{\frac{h^2}{2}\lambda\nabla^2\Omega(x)d_1}_{(E1)}, \tag{33}$$

and

$$\begin{aligned}(A3)+(D2) &= \frac{1}{2}D^3f(x)[\tilde{x}-x]^2+\frac{h}{2}\left(\nabla^2 f(x_1)-\nabla^2 f(x_2)\right)d_2\\ &= \frac{h^2}{2}D^3f(x)\left[\frac{1}{2}(d_1+d_2)\right]^2+\frac{h}{2}D^3f(x)[x_1-x_2,d_2]+(R1)\\ &= \frac{h^2}{2}D^3f(x)\left[\frac{1}{2}(d_1+d_2)\right]^2-\frac{h^2}{2}D^3f(x)[d_1,d_2]+(R1)\\ &= \underbrace{\frac{h^2}{2}D^3f(x)\left[\frac{1}{2}(d_1-d_2)\right]^2}_{(R2)}+(R1),\end{aligned}$$

where

$$\begin{aligned}\|(R1)\| &= \left\|\frac{h}{2}\left(\nabla^2 f(x_1)-\nabla^2 f(x_2)\right)d_2-\frac{h}{2}D^3f(x)[x_1-x_2,d_2]\right\|\\ &\le \frac{h}{2}\cdot\frac{L_{f,3}}{2}\|x_1-x_2\|^2\|d_2\|=\frac{h^3}{4}L_{f,3}\|d_1\|^2\|d_2\|.\end{aligned}$$

We further split $(D3)$ into

$$\begin{aligned}(D3) &= \frac{h}{2}(\lambda_1\nabla^2\Omega(x_1)-\lambda_2\nabla^2\Omega(x_2))d_2\\ &= \underbrace{\frac{h}{2}\lambda_2\left(\nabla^2\Omega(x_1)-\nabla^2\Omega(x_2)\right)d_2}_{(E2)}+\underbrace{\frac{h}{2}(\lambda_1-\lambda_2)\nabla^2\Omega(x_1)d_2}_{(E3)},\end{aligned}$$

and then rearrange them by

$$\begin{aligned}(B4)+(E2) &= \tilde{\lambda}\cdot\frac{1}{2}D^3\Omega(x)[\tilde{x}-x]^2+\frac{h}{2}\lambda_2\left(\nabla^2\Omega(x_1)-\nabla^2\Omega(x_2)\right)d_2\\ &= \tilde{\lambda}\cdot\frac{1}{2}D^3\Omega(x)[\tilde{x}-x]^2+\frac{h}{2}\lambda_2D^3\Omega(x_1)[x_1-x_2,d_2]+(R5)\\ &= \underbrace{\frac{h^2}{2}(\tilde{\lambda}-\lambda_2)D^3\Omega(x)\left[\frac{d_1+d_2}{2}\right]^2}_{(R3)}+\underbrace{\frac{h^2}{2}\lambda_2D^3\Omega(x_1)\left[\frac{d_1-d_2}{2}\right]^2}_{(R4)}+(R5),\end{aligned}$$

where

$$\begin{aligned}\|(R5)\| &= \left\|\frac{h}{2}\lambda_2\left(\nabla^2\Omega(x_1)-\nabla^2\Omega(x_2)\right)d_2-\frac{h}{2}\lambda_2D^3\Omega(x)[x_1-x_2,d_2]\right\|\\ &\le \frac{h}{2}\lambda_2\cdot\frac{L_{\Omega,3}}{2}\|x_1-x_2\|^2\|d_2\|=\frac{h^3}{4}\lambda_2L_{\Omega,3}\|d_1\|^2\|d_2\|,\end{aligned}$$

and

$$\begin{aligned}(B3)+(E1)+(E3) &= (\tilde{\lambda}-\lambda)\nabla^2\Omega(x)(\tilde{x}-x)+\frac{h^2}{2}\lambda\nabla^2\Omega(x)d_1+\frac{h}{2}(\lambda_1-\lambda_2)\nabla^2\Omega(x_1)d_2\\ &= \left(1-h+\frac{h^2}{2}-1\right)\lambda\nabla^2\Omega(x)h\cdot\frac{d_1+d_2}{2}+\frac{h^2}{2}\lambda\nabla^2\Omega(x)d_1\\ &\quad+\frac{h}{2}\left(1-1+h-h^2\right)\nabla^2\Omega(x)d_2\\ &= \underbrace{\frac{h^3}{4}\lambda\nabla^2\Omega(x)(d_1-d_2)}_{(R6)}.\end{aligned}$$

Hence, it holds that

$$\begin{aligned}
\|(R)\| &\le \|(RA)\| + \|(RB)\| + \|(RD)\| + \|(R1)\| + \|(R2)\| \\
&\quad + \|(R3)\| + \|(R4)\| + \|(R5)\| + \|(R6)\| \\
&\le h^3 \cdot \frac{L_{f,3}}{6} \left\| \frac{d_1+d_2}{2} \right\|^3 + h^3 \cdot \frac{\tilde{\lambda} L_{\Omega,3}}{6} \left\| \frac{d_1+d_2}{2} \right\|^3 \\
&\quad + h^3 \cdot \frac{L_{f,2}}{6} \|d_1\|^2 + h^3 \cdot \frac{L_{f,3}}{4} \|d_1\|^2 \|d_2\| \\
&\quad + h^2 \cdot \frac{L_{\Omega,2}}{8} \|d_1 - d_2\|^2 + h^4 \cdot \frac{\lambda L_{\Omega,2}}{4} \left\| \frac{d_1+d_2}{2} \right\|^2 + h^2 \cdot \frac{\lambda_2 L_{\Omega,2}}{8} \|d_1 - dy_2\|^2 \\
&\quad + h^3 \cdot \frac{\lambda_2 L_{\Omega,3}}{4} \|d_1\|^2 \|d_2\| + h^3 \cdot \frac{\lambda L_{\Omega,1}}{4} \|d_1 - d_2\| \\
&\le h^3 \cdot \frac{L}{3} \|d_1\|^3 + h^3 \cdot \frac{L\lambda}{3} \|d_1\|^3 + h^3 \cdot \frac{L}{6} \|d_1\|^2 + h^3 \cdot \frac{L}{2} \|d_1\|^3 + h^2 \cdot \frac{L}{8} \|d_1 - d_2\|^2 \\
&\quad + h^4 \cdot \frac{\lambda L}{2} \|d_1\|^2 + h^2 \cdot \frac{\lambda L}{8} \|d_1 - d_2\|^2 + h^3 \cdot \frac{\lambda L}{2} \|d_1\|^3 + h^3 \cdot \frac{\lambda L}{4} \|d_1 - d_2\| \\
&\le h^3 L(1+\lambda) \left(\|d_1\|^3 + \|d_1\|^2 + \|d_1\| \right) + h^2 \cdot \frac{L(1+\lambda)}{8} \|d_1 - d_2\|^2 .
\end{aligned} \tag{34}$$

From Lemma 7, it holds that $\|d_1 - d_2\| \le \frac{h}{\tilde{\mu}_x} \cdot 2L(1+\lambda) \left(\|d_1\| + \|d_1\|^2 \right)$. Apply this inequality into (34), we can further get

$$\begin{aligned}
\|d_1 - d_2\|^2 &\le \left(\frac{h}{\tilde{\mu}_x} \cdot 2L(1+\lambda) \left(\|d_1\| + \|d_1\|^2 \right) \right)^2 \\
&\le \frac{8h^2 L^2}{\tilde{\mu}_x^2} \left((1+\lambda)^2 \|d_1\|^2 + (1+\lambda)^2 \|d_1\|^4 \right) \\
&\le \frac{8h^2 L^2}{\tilde{\mu}_x^2} \left((1+\lambda)^2 \|d_1\|^2 + (1+\lambda)^2 \|d_1\|^4 \right)
\end{aligned} \tag{35}$$

Also, by applying (26) into (35) we have

$$\|(R)\| \le h^3 \cdot 3L(1+G)^3 + h^4 \cdot \frac{2L^3(1+G)^4}{\tilde{\mu}_x^2}.$$

□

Theorem 7. *Suppose μ-strongly convex function f and 1-strongly convex function $\Omega(\cdot)$ satisfy Assumption 2. Consider the set of interest $[\lambda_{\min}, \lambda_{\max}]$. Let x_0 be the initialization at $\lambda_0 = \lambda_{\max}$ and $r_0 = \|\nabla f(x_0) + \lambda_0 \nabla \Omega(x_0)\|$ be the initialization residual. Then the update rule $(x_{k+1}, \lambda_{k+1}) = T(x_k, \lambda_k; h)$ when step-size h satisfies*

$$h \le \min \left\{ 0.2, \frac{\tilde{\mu}_{\min}}{8L(1+G)}, \frac{\epsilon^{1/2}}{6L^{1/2}(1+G)^{3/2}}, \frac{\tilde{\mu}_{\min}^{2/3} \epsilon^{1/3}}{5L(1+G)^{4/3}} \right\}, \tag{36}$$

we have for all k such that $\lambda_k \ge \lambda_{\min}$ it holds that

$$r_k := \|\nabla f(x_k) + \lambda_k \nabla \Omega(x_k)\| \le \epsilon. \tag{37}$$

Proof. Prove by induction. Suppose $r_k \le \epsilon$, then by Theorem 6, it holds that

$$r_{k+1} \le \left(1 - h + \frac{h^2}{2} \right) r_k + h^3 \cdot 3L(1+G)^3 + h^4 \cdot \frac{2L^3(1+G)^4}{\tilde{\mu}_x^2}.$$

Also, using the inequalities in (36), we have

$$\frac{h^2}{2} \le 0.1h, \quad h^2 \cdot 3L(1+G)^3 \le 0.5\epsilon, \quad h^3 \cdot \frac{2L^3(1+G)^4}{\tilde{\mu}_x^2} \le 0.4\epsilon.$$

Hence, it holds that $r_{k+1} \le (1 - 0.9h) r_k + 0.5h\epsilon + 0.4h\epsilon \le \epsilon$. □

Theorem 6 shows that by setting $h \sim O(\epsilon^{1/2})$ the trapezoid guarantees residual at each near optimal solution is smaller than ϵ. For all other $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, we implement linear interpolation to approximate the corresponding near optimal solution. We now complete the formal proof of Theorem 2, which provides the complexity analysis of Algorithm 1 with update scheme 4.

Proof of Theorem 2. Choose step-size h to be the smaller root of $1 - h + \frac{h^2}{2} = \left(\frac{\lambda_{\min}}{\lambda_{\max}}\right)^{1/K}$. Since $h - \frac{h^2}{2} \leq \frac{1}{K} \log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)$, step-size h satisfies the conditions in (36). It implies that for all k such that $\lambda_k \leq \lambda_{\min}$ it holds that $r_k \leq \epsilon$.

Then consider interpolation error. Suppose $\lambda \in [\lambda_{k+1}, \lambda_k]$ and hence $\hat{x}(\lambda) = \alpha x_k + (1-\alpha) x_{k+1}$ where $\alpha = \frac{\lambda - \lambda_{k+1}}{\lambda_k - \lambda_{k+1}}$. Applying results in Theorem 4 it holds that

$$\begin{aligned}
\|f(\hat{x}(\lambda)) + \lambda \hat{x}(\lambda)\| &\leq \alpha r_k + (1-\alpha) r_{k+1} + h^2 \cdot \frac{L_{f,2}}{8} \|d_j\|^2 \\
&\quad + h^2 \cdot \left(\frac{\lambda L_{\Omega,2}}{8} \|d_j\|^2 + \frac{|c_j| L_{\Omega,1}}{4} \|d_j\| \right) \\
&\leq \epsilon + h^2 \cdot L(1+\lambda) \left(\frac{1}{8} \|d_j\|^2 + \frac{1}{4} \|d_j\| \right) \\
&\leq \epsilon + h^2 \cdot L(1+G)^2 \leq 2\epsilon.
\end{aligned}$$

□

A.3 Proof of Proposition 1

We prove the following two lemma instead, which are stronger versions of Corollary 1.

Lemma 8 (Analysis of Approximate Euler Update)**.** *Under Assumption 1, discretization* (3) *with δ-accurate direction, it holds that*

$$r_{k+1} \leq \left(1 + h \cdot \frac{c_k}{\lambda_k}\right) r_k + h^2 \cdot \frac{L_{f,2} + \lambda_{k+1} L_{\Omega,2}}{2} \|d_k\|^2 + h \|\delta_k\|. \tag{38}$$

Furthermore, if we set $\lambda_{k+1} = (1-h)\lambda$ and $\|\delta_k\| \leq \delta$ for some scalar $\delta > 0$, it holds that

$$\begin{aligned}
r_k &\leq \frac{\lambda_k}{\lambda_0} \cdot r_0 + 2h \left(\frac{L_{f,2}}{\tilde{\mu}} + L_{\Omega,2} \right) R_f + \sum_{i=1}^{k} h(1-h)^{k-i} \|\delta_k\| \\
&\leq \frac{\lambda_k}{\lambda_0} \cdot r_0 + 2h \left(\frac{L_{f,2}}{\tilde{\mu}} + L_{\Omega,2} \right) R_f + \delta.
\end{aligned} \tag{39}$$

Proof. The proof of (38) is similar with the proof in Lemma 2. When d_k is an inexact solution, the right hand side of (8) becomes $\left(1 + h \cdot \frac{c_k}{\lambda_k}\right)(\lambda \nabla \Omega(x_k) + \nabla f(x)) + h\delta_k$. Also, applying (38) to Proposition 3 implies the result in (39). □

Lemma 9 (Robustness of trapezoid)**.** *Suppose $x \in S_{x_0}, \lambda \in [\lambda_{\min}, \lambda_{\max}], h > 0$ and $(\tilde{x}, \tilde{\lambda}) = T(x, \lambda; h, \delta)$. Let $r = \|\nabla f(x) + \lambda \nabla \Omega(x)\|$ satisfying that $r \leq \tilde{\mu}_x$. Then for all $x \in \mathbb{R}^n$ and $\lambda > 0$, it holds that*

$$\tilde{r} := \left\|\nabla F_{\tilde{\lambda}}(\tilde{x})\right\| \leq \frac{\tilde{\lambda}}{\lambda} \|\nabla F_\lambda(x)\| + h^3 \cdot 3L(c+G)^3 + h^4 \cdot \frac{2L^3(c+G)^4}{\tilde{\mu}_x^2} + \frac{h}{2} \|\delta_1 - \delta_2\| + \frac{h^2}{2} \|\delta_1\|, \tag{40}$$

where $c = \|\delta_1\| / \tilde{\mu}_x + 1$. Furthermore, suppose the sequence $\{(x_k, \lambda_k)\}_{k=0}^K$ is generated by $(x_{k+1}, \lambda_{k+1}) = T(x_k, \lambda_k; h, \delta_k)$ and $\|\delta_{k,\cdot}\| \leq \delta$ for some scalar $\delta > 0$, then it holds that

$$\max_{k \in [K]} \{r_k\} \leq h^2 \cdot 3L(c+G)^3 + h^3 \cdot \frac{2L^3(c+G)^4}{\tilde{\mu}_x^2} + \frac{2+h}{2} \delta. \tag{41}$$

Proof. We will follow the idea in Lemma 6, 7 and Theorem 6. Recall the result in Lemma 6, since $\tilde{H}_1 d_1 = -\nabla f(x_1) + \delta_1$, it holds that $(1+\lambda)\|d_1\| \leq 2(G+1+\|\delta_1\|/\tilde{\mu})$. Also the result in Lemma 7 becomes

$$\left\|\tilde{H}_1(d_2-d_1) - \nabla^2 f(x_1)(x_1-x_2) - (\tilde{H}_1-\tilde{H}_2)d_2 + \delta_1 - \delta_2\right\| \leq \frac{L_{f,2}}{2}\|x_1-x_2\|^2,$$

where $\tilde{H}_1 = \nabla^2 f(x_1) + \lambda_1 \nabla^2 \Omega(x_1)$ and $\tilde{H}_2 = \nabla^2 f(x_2) + \lambda_2 \nabla^2 \Omega(x_2)$. Now we modify the proof of Theorem 6 to get (40). Then the right hand side of (31) becomes $-h\delta_1$. Also, the right hand side of (32) becomes $\frac{h}{2}(\delta_1 - \delta_2)$ and the right hand side of (33) becomes $\frac{h^2}{2}\lambda\nabla^2\Omega(x)d_1 + \frac{h^2}{2}\delta_1$. □

A.4 Proof of Corollary 1

At iteration k, second-oracle conjugate gradient variant of update scheme (3) requires an approximate solution $\hat{d}_k$ satisfying $\|H_k\hat{d}_k + g_k\|_2 \leq \delta$ where $H_k := \nabla^2 f(x_k) + \lambda_{k+1}\nabla^2\Omega(x_k)$ and $g_k := \nabla f(x_k)$. At each iteration of conjugate gradient method, we need to compute $H_k \cdot p_{k,t}$ where $\{p_{k,t}\}$ are the orthogonal basis of matrix H_k. Hence, exact one Hessian-vector product oracle will be called at each iteration to compute $\hat{d}_k$. Let $\{y_{k,t}\}$ denote the sequence generated by conjugate gradient method with $y_{k,0} = \hat{d}_{k-1}$. Existing results of conjugate gradient method guarantees that

$$\|H_k y_{k,t} + g_k\|_2 \leq 2\sqrt{\kappa_k}\left(1 - \frac{2}{\sqrt{\kappa_k}+1}\right)^t \|H_k y_{k,0} + g_k\|_2,$$

where κ_k is the condition number of H_k. Since the initial guess $y_0 = \hat{d}_{k-1}$ which is the approximate direction at last iteration, we have $\|H_{k-1}\hat{d}_{k-1} + g_{k-1}\| \leq \epsilon$. Then the initial guess guarantees that

$$\begin{aligned}
\|H_k\hat{d}_{k-1} + g_k\|_2 &\leq \|H_{k-1}\hat{d}_{k-1} + g_{k-1}\|_2 + \|H_k\hat{d}_{k-1} + g_k - H_{k-1}\hat{d}_{k-1} - g_{k-1}\| \\
&\leq \epsilon + h\left((L_{f,2} + \lambda_k L_{\Omega,2})\|\hat{d}_{k-1}\|^2 + (\lambda_k L_{\Omega,1} + L_{f,1})\|\hat{d}_{k-1}\|\right) \\
&\leq \epsilon + 2hL(2+G)^2.
\end{aligned}$$

We apply the step-size $h = \frac{\tilde{\mu}\epsilon}{LR_f}$ from Theorem 1. Let N_k denote the number of inner iteration at iteration k and we have

$$N_k \leq \frac{\sqrt{\kappa_k}+1}{2}\log\left(\frac{2\sqrt{\kappa_k}\|H_k y_{k,0} + g_k\|_2}{\epsilon}\right) \sim \tilde{\mathcal{O}}(\sqrt{\kappa_k}).$$

Here we hide logarithmic terms in the $\tilde{\mathcal{O}}(\cdot)$ notation. Since κ_k is the condition number of the matrix $\nabla^2 f(x_k) + \lambda_{k+1}\nabla^2\Omega(x_k)$, we have

$$\kappa_k \leq \max\left\{\kappa\left(\nabla^2 f(x_k)\right), \kappa\left(\nabla^2\Omega(x_k)\right)\right\}.$$

Let κ denote the larger value between the condition number of $\nabla^2 f(\cdot)$ and $\nabla^2\Omega(\cdot)$, we have $\kappa_k \leq \kappa$ for all k, where $\kappa := \frac{L}{\tilde{\mu}}$ Hence, we pay another $\tilde{\mathcal{O}}(\sqrt{\kappa})$ factor to apply the Hessian-free variant via the conjugate gradient sub-routine. The argument of second-oracle conjugate gradient variant of update scheme (4) is similar.