**Title**
New Results for Online and Offline Stochastic Optimization and Decision-Making



**Author**
Liu, Heyuan

**Publication Date**
2022

Peer reviewed|Thesis/dissertation

# New Results for Online and Offline Stochastic Optimization and Decision-Making

by

Heyuan Liu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Paul Grigas, Chair
Associate Professor Anil Aswani
Associate Adjunct Professor Michael Mahoney
Assistant Professor Zeyu Zheng

Summer 2022

New Results for Online and Offline Stochastic Optimization and Decision-Making

Abstract

New Results for Online and Offline Stochastic Optimization and Decision-Making

by

Heyuan Liu

Doctor of Philosophy in Engineering – Industrial Engineering and Operations Research

University of California, Berkeley

Assistant Professor Paul Grigas, Chair

This dissertation presents several contributions at the interface of methods for convex optimization problems and decision-making problems in both online and offline settings.

The first part of the dissertation focuses on new optimization methods for computing an approximate solution path for parameterized optimization problems. We develop and analyze several different second-order algorithms for computing a near-optimal solution path of a convex parametric optimization problem with smooth Hessian. Our algorithms are inspired by a differential equation perspective on the parametric solution path and do not rely on the specific structure of the objective function. We present computational guarantees that bound the oracle complexity to achieve a near-optimal solution path under different sets of smoothness assumptions. Under the assumptions, the results are an improvement over the best-known results of the grid search methods. We also develop second-order conjugate gradient variants which avoid exact computation of Hessian and solving linear equations. We present computation results that demonstrate the effectiveness of our methods over the grid search methods on both real and synthetic datasets. On large-scale problems, we demonstrate significant speedups of the second-order conjugate variants as compared to the standard versions of our methods.

The second part of the dissertation focuses on the statistical properties of the recently introduced surrogate "SPO+" loss function in the "Smart Predict-then-Optimize (SPO)" framework. We greatly expand upon the consistency results for the surrogate loss in previous literature. We develop risk bounds and uniform calibration results for the surrogate loss relative to the original loss, which provide a quantitative way to transfer the excess surrogate risk to excess true risk. By combining our risk bounds with generalization bounds, we show that the empirical minimizer of the surrogate loss achieves low excess true risk with high probability. We first demonstrate these results in the case when the feasible region of the underlying optimization problem is a polyhedron, and then we show that the results can be strengthened substantially when the feasible region is a level set of a strongly convex

function. We perform experiments to empirically demonstrate the strength of the SPO+ surrogate, as compared to standard $\ell_1$ and squared $\ell_2$ prediction error losses, on portfolio allocation and cost-sensitive multi-class classification problems.

The third part of the dissertation focuses on the online contextual decision-making problem with resource constraints. We propose an algorithm that mixes a prediction step based on the SPO method with a dual update step based on mirror descent. We prove regret bounds and demonstrate that the overall convergence rate of our method depends on the $\mathcal{O}(T^{-1/2})$ convergence of online mirror descent as well as risk bounds of the surrogate loss function used to learn the prediction model. Our algorithm and regret bounds apply to a general convex feasible region for the resource constraints, including both hard and soft resource constraint cases, and they apply to a wide class of prediction models in contrast to the traditional settings of linear contextual models or finite policy spaces. We also conduct numerical experiments to empirically demonstrate the strength of our proposed SPO-type methods, as compared to traditional prediction-error-only methods, on multi-dimensional knapsack and longest path instances.

To Jianwen, Yilin, and Qi

# Contents

# List of Figures

# Acknowledgments

First and foremost, I would like to thank my advisor Professor Paul Grigas. I am immensely fortunate to be advised by Paul, who is a patient professor, a supportive advisor, and a gracious friend. The past five years at Berkeley have been unforgettable experiences for me, and working with Paul, has been the best part of it. Paul has always been inspiring, encouraging, and enthusiastic, and he is always incredibly helpful with guidance, advice, and assistance. Every time we met, Paul always had novel ideas, precise feedback, and valuable advice, and also encouraged me to explore further our research projects. I am so grateful, and could not even imagine before my graduate study, the very unique and overwhelmingly positive experience while learning from and working with Paul, and I hope I could become a person like him.

Secondly, I would like to thank my dissertation committee member, Professor Anil Aswani, Professor Michael Mahoney, Professor Max Shen, and Professor Zeyu Zheng for offering insightful feedback and advice during my graduate study. I am also grateful to other faculty members in IEOR and College of Engineering for offering a variety of excellent courses, which intrigue my research interests and provide useful tools. Same great thanks to IEOR staff and Berkeley research computing for their daily support and continuous help.

Next, I would like to thank my colleagues and friends, from both inside and outside of IEOR, especially Eric Bertelli, Caleb Bugg, Haoyang Cao, Junyu Cao, Yuhao Ding, Han Feng, Pedro Hespanhol, Hansheng Jiang, Yusuke Kikuchi, Anran Hu, Feng Ji, Tianyi Lin, Mo Liu, Alfonso Lobos, Igor Molybog, Pelagie Elimbi Moudio, Matt Olfat, Meng Qi, Xu Rao, Quico Spaen, Yu Tong, Mark Velednitsky, Renyuan Xu, Nan Yang, Zitong Yang, Junzi Zhang, and Ruijie Zhou. The conversations and laughter with you are always enjoyable. I owe gratitude to you who supported me and made the journey so memorable and fun.

I would also like to thank some faculty members during my undergraduate study, Professor Chenxu Li, Professor Tiejun Li, Professor Jannik Matuschke, Professor Andreas S. Schulz, and Professor Zaiwen Wen, just to name a few. The undergraduate program at the School of Mathematical Sciences, Peking University was incredible, which provided sufficient mathematical knowledge and tools, and intrigued my initial interests in operations research as well. I also appreciate the internship opportunities at Theorem and Two Sigma. My managers Federico Gonzalez and Hari Adireddy were always kind and helpful. A big thanks to my intern buddies Yu Wang and Bumeng Zhuo who provided a warm working environment.

Finally, my deeply grateful acknowledgment goes to my family for their endless love and support. My parents and role models, Jianwen and Yilin, have provided a tremendous amount of resources, both financially and mentally in my pursuit of personal and academic goals. They have been supportive of any decisions I made, and always available for suggestions when I was in doubt. "Honesty and integrity" are the words how my parents describe me and I will always remember that. Thanks to my spouse, Qi, without whom my life will be incomplete. We supported, accompanied, and encouraged each other during the most difficult times of the COVID-19 epidemic. All of my achievements are due to the unconditional love and support of my family, and I dedicate this dissertation to them.

# Chapter 1

# Introduction

The recent two decades witnessed the substantial developments of both optimization and learning methods thanks to the advance in computing power and data storage. In the operations research community, it is crucial to fully leverage modern techniques and massive amount of data to develop predictive models, design efficient algorithms, and finally, make good decisions, especially when the decision-making problems are associated with uncertainty. This dissertation presents new algorithms and theoretical guarantees for different stochastic optimization and decision-making problems, in both online and offline settings.

In Chapter 2, the main focus is the parametric optimization problem, where the objective function depends on one parameter. In many applications of interest, it is necessary to solve not just a single optimization problem but an entire collection of related problems as a function of the parameter. There are several strong motivations to design algorithms for the parametric problems, including but not limited to: *(i)* the need to solve problems arising in application areas like regularized regression with cross-validation, and *(ii)* the need to address multi-objective optimization, for instance, finding the Pareto frontier of a two-objective optimization problem. An important set of problems in practice and a popular line of research involves computing the solution path of regularized machine learning problems, including the LASSO and the SVM problem. In these works, algorithms are designed to compute the exact piecewise linear solution paths. In contrast, we consider general convex objective functions and design efficient algorithms to compute the approximate solution paths. We provide a novel perspective to analyze the solution path by deriving an ordinary differential equation whose solution is the exact solution path. This understanding enables us to design more efficient algorithms which are adaptive to the smoothness of objective functions, and present computational guarantees respectively. We modify the standard update schemes in numerical ordinary differential equations and develop non-asymptotic complexity bounds. We then incorporate linear interpolation, which was often missing either in practical algorithms or in the lower bounds complexity analysis, to generate nearly-optimal solutions for the entire parameter interval of interest. Our complexity analysis measures the number of operations, such as Hessian evaluations, rather than the number of sub-problems that are required to be solved, such as individual optimization problems and numerical differential equations

as has been studied in recent works. In large-scale problems, to avoid computing Hessian and/or solving linear systems, which are required in the aforementioned exact algorithms, we consider second-order conjugate gradient type methods. By leveraging the theoretical analysis in the presence of inexact directions, we provide the computational guarantees of the new variants, which have the same order of the desired accuracy as the results for the exact algorithms.

In Chapter 3, the main focus is the offline contextual decision-making problems, where one first predicts the unknown parameters in a decision-making problem and then plugs in the prediction before solving the problem. This framework has a wide variety of applications, including navigation problems, wherein the actual travelling time on each edge is unavailable when the routes need to be recommended, and portfolio allocation problems, wherein the expected return is unavailable. Most practical methods utilize the contextual information, for example, time of day, weather information, and, financial and business news headlines, to infer the actual parameter and reduce the uncertainty. Ultimately, the goal is to produce a high-quality prediction model that leads to a good decision when implemented. A natural loss function in this setting is defined by measuring the decision error induced by the predicted parameters, which was named the Smart Predict-then-Optimize (SPO) loss by [29]. The authors also introduced the surrogate SPO+ loss due to the non-convexity and non-continuity of the original SPO loss and provide the Fisher consistency under mild conditions. We extend the analysis by considering the excess risk bounds of the surrogate SPO+ loss function, which answer the following question: to what tolerance $\delta$ should the excess surrogate risk be reduced to in order to ensure that the excess SPO risk is at most $\epsilon$? By making use of uniform calibration, we developed risk bounds in the two cases depending on the structure of the feasible region of the optimization problem: *(i)* the case of a bounded polyhedron, and *(ii)* the case of a level set of a smooth and strongly convex function. As a consequence of our analysis, we can leverage generalization guarantees for the SPO+ loss to obtain the first sample complexity bounds, with respect to the SPO risk, for the SPO+ surrogate under the two cases we consider. We also provide a faster convergence rate when the distribution satisfies some certain low near-degeneracy conditions.

In Chapter 4, the problem of interest is the online contextual decision-making problems. The final goal of the decision-maker is to maximize the summation of the reward and the utility from resource consumption, while satisfying the resource constraints. We develop a new framework for integrating decision-focused learning methods, using predict-then-optimize losses, into the online decision-making task. The main difference from the previous chapter is that there is a trade-off between immediate rewards and rewards received at a later time, where the trade-off exists since each decision that is made consumes some of a limited amount of resources. To address the resource consumption, we apply the technique of introducing dual variables and using primal-dual methods. In the decision step, we need to predict the reward and consumption to solve a linear optimization problem with a known feasible region to make a decision. Due to the linear structure of the underlying optimization problem in our meta-procedure, we can apply the aforementioned SPO loss function and its surrogate losses. As such, at each time period, we update a set of dual variables using the method of

online mirror descent and then we update the prediction model by minimizing a surrogate of the SPO loss on a dataset constructed by combining past observations with the current dual variables. A critical part of our contribution involves bridging convergence theory for primal-dual online methods with learning theory in the predict-then-optimize setting. In particular, we prove regret bounds for our overall algorithm that combine the $\mathcal{O}(T^{-1/2})$ convergence of online mirror descent with the convergence of the learning process, the rate of which depends on which surrogate loss function is used. Our algorithm and analysis are no longer limited to the previously studied linear context or finite policy assumptions, and more complex machine learning models, such as random forests and neural networks, may be used. Our bounds hold in both hard and soft resource constraint cases, and we extend prior results using standard upper bound consumption constraints on each resource to arbitrary convex consumption constraints.

# Chapter 2

# New Methods for Solution Path Optimization via Differential Equations

## 2.1 Introduction

In many applications of interest, it is necessary to solve not just a single optimization problem but an entire collection of related problems. In these settings, some or all of the objects involved in defining the objective function or constraints of an optimization problem depend on one or more parameters, and we would like to solve the problem as a function of these parameters. Generally, a *parametric optimization problem* can be written as:

$$P(\lambda): \quad \min_{x \in S(\lambda)} F(x, \lambda), \tag{2.1}$$

where $\lambda$ belongs to the set of interest $\Lambda \subseteq \mathbb{R}^m$, and the feasible sets satisfy $S(\lambda) \subseteq \mathbb{R}^p$. There are many problems of interest that are formulated as parametric optimization problems of the form (2.1). Subsequently, as indicated by [38], there are several strong motivations to design algorithms for (2.1), including but not limited to: *(i)* the need to solve problems arising in application areas like regularized regression with cross-validation (see, e.g., [69]) and model predictive control (see, e.g., [33]), *(ii)* as a building block for developing globally convergent algorithms by the approach of path-following as is done in interior-point methods (see, e.g., [65]), and *(iii)* the need to address multi-objective optimization, for instance, finding the Pareto frontier of a two-objective optimization problem. Depending upon the assumptions made, the goal may be to find global/local optimal solutions or Karush–Kuhn–Tucker (KKT) points of problem $P(\lambda)$ for $\lambda \in \Lambda$.

In the rest of the paper, we will focus on a more specific problem, in which we assume: *(i)* the dependence on $\lambda$ is linear, that is, $F(x, \lambda)$ can be written as $f(x) + \lambda \cdot \Omega(x)$, *(ii)* both $f(\cdot)$ and $\Omega(\cdot)$ are convex functions with certain properties, and *(iii)* the feasible set $S(\lambda)$ is the entire vector space $\mathbb{R}^p$ for all $\lambda \in \Lambda$. That is, we focus on the parametric optimization

problem:

$$P(\lambda): \quad F_\lambda^* := \min_{x\in\mathbb{R}^p} \{F_\lambda(x) := f(x) + \lambda \cdot \Omega(x)\}, \tag{2.2}$$

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 a $\mu$-strongly convex for some $\mu \geq 0$ and $\Omega(\cdot)$ is $\sigma$-strongly convex for some $\sigma > 0$, both with respect to the $\ell_2$-norm (denoted by $\|\cdot\|$ herein). For any $\lambda > 0$, let

$$x(\lambda) := \arg\min_{x\in\mathbb{R}^p} F_\lambda(x) \tag{2.3}$$

denote the unique optimal solution of $P(\lambda)$ defined in (2.2). We are interested in the problem of (approximately) computing the set of optimal solutions $\{x(\lambda) : \lambda \in \Lambda\}$ where $\Lambda = [\lambda_{\min}, \lambda_{\max}]$ is the set of interest for some $0 < \lambda_{\min} < \lambda_{\max}$, and we also refer to this set of solutions as the *(exact) solution path*. An important set of problems in practice and a popular line of research involves computing the solution path of regularized machine learning problems, including the LASSO as in [27, 69] and the SVM problem as in [39]. In these works, algorithms are designed to compute the exact piecewise linear solution paths. Also, in the context of interior-point method for constrained convex optimization (see, for instance, [65] and [74]), $f(\cdot)$ represents the objective function and $\Omega(\cdot)$ represents the barrier function induced by the constraints of origin problem. Note that this application requires a slightly more general version of problem (2.2) where $\Omega(\cdot): \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$. The interior-point method starts with the problem $P(\lambda)$ with a moderately large $\lambda_0$ and terminates when $\lambda_k < \delta$ for some small enough positive threshold $\delta$. Recently, there has been growing interest in developing algorithms for computing an approximate solution path of a generic problem like (2.3). [75] consider applying exact Newton steps on pre-specified grids, and [63] consider adaptive methods to discretize the interval $[\lambda_{\min}, \lambda_{\max}]$, for example. These grid search type methods, which discretize the interval $[\lambda_{\min}, \lambda_{\max}]$ and subsequently solve a sequence of individual optimization problems take a very black-box approach. A natural question is: can we "open the black-box" by developing a methodology that better exploits the structure of the solution path? We answer this question positively by introducing a differential equation perspective to analyze the solution path, which enables us to better reveal and exploit the underlying structure of the solution path. This deeper understanding enable us to build more efficient algorithms and present improved computational guarantees.

In particular, we derive an ordinary differential equation with an initial condition whose solution is the exact solution path of (2.2). The dynamics of the ODE that we derive resemble, but are distinct from, the dynamics of a path-wise version of Newton's method. Based on the ODE, we propose efficient algorithms to generate *approximate solution paths* $\hat{x}(\lambda) : \lambda \in [\lambda_{\min}, \lambda_{\max}] \to \mathbb{R}^p$ and we provide the corresponding complexity analysis. The metric we consider is the $\ell_2$-norm of the gradient of the regularized problem, namely $\|\nabla F_\lambda(\hat{x}(\lambda))\|_2$, and we use the largest norm along the approximate path $\sup_{\lambda\in\Lambda} \|\nabla F_\lambda(\hat{x}(\lambda))\|_2$ to represent the accuracy of an approximate path $\hat{x}(\lambda)$ (as formally defined in Definition 2.3.1). To analyze the computational cost of our proposed algorithms, we consider the oracle complexity – either in terms of full Hessian or Hessian-vector product/gradient evaluations – to obtain an $\epsilon$-accurate solution path. Note that considering the oracle complexity is

is contrast to other works that consider the number of individual optimization problems that need to be solved (see, for example, [35, 63]), as well as the number of ordinary differential equations that need to be solved (see, for example, [90]).

### 2.1.1 Contributions

The first set of contributions of this paper concern the perspective of the solution path of (2.2) from an ordinary differential equation point of view. We derive an ordinary differential equation with an initial condition whose solution is the solution path of (2.2), based on the first-order optimality conditions of (2.2). With this observation, we propose a novel and efficient way to approximate the entire solution path. Our derivation does not rely on the special structure of the optimization problem, like existing results in the solution path of LASSO or SVM problems, and holds for general objective functions.

The second set of contributions of this paper concern the design of efficient algorithms and the corresponding oracle complexity analysis. Classical error analysis of numerical ordinary differential equation methods in [34, 78] provide only asymptotic results, and the global error has an exponential dependency on the Lipschitz constant and the length of the time period. In contrast, we design new update schemes to compute an approximate solution path and develop non-asymptotic complexity bounds. In particular, we apply a semi-implicit Euler method on the ordinary differential equation in order to compute the approximate optimal solutions under a finite set of penalty coefficients. Then, we incorporate linear interpolation, which was usually missing either in practice or in the lower bound complexity analysis, to generate nearly-optimal solutions under other penalty coefficients within the range of parameter values of interest. The two-step algorithm guarantees an $\epsilon$-accurate solution path within at most $\mathcal{O}(\frac{1}{\epsilon})$ gradient and Hessian evaluations as well as linear equation system solves. When the objective function has higher-order smoothness properties, we modify the traditional trapezoid method in numerical differential equations and design a new update scheme, which guarantees an $\epsilon$-path within at most $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ Hessian evaluations. It is important to emphasize that the complexity results in this paper are in terms of the number of operations (for example, Hessian evaluations), rather than the number of sub-problems that need to be solved (for example, solving a single numerical ODE, or individual optimization problems) as has been studied in prior work [35, 90]. We also provide a detailed computational evaluation of our algorithms and existing methods, including several experiments on synthetic data, the breast cancer dataset [26], and the leukemia dataset [37].

The third set of contributions of the paper concern second-order conjugate gradient type methods and computational guarantees in the presence of inexact gradient and Hessian oracles as well as approximate linear equation solvers. When the dimension of the problem is high, computing the Hessian and/or solving linear systems becomes a computational bottleneck. To avoid this, one would like an algorithm that only requires approximate directions at each iteration. We first consider the case when (absolute) numerical error incurred in the calculation of a new direction $d_k$ is bounded by some $\delta_k > 0$. We show that our algorithms are robust to numerical error in the sense that the additional errors of

inexact directions does not accumulate and does not depend on the condition number. We extend the complexity analysis to the case when the numerical error $\delta_k$ has a uniform upper bound $\alpha\epsilon$ for $\alpha \in (0, 1)$ and show that the Euler method maintains $\mathcal{O}(\frac{1}{\epsilon})$ complexity, and the trapezoid method maintains $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ complexity when there is higher-order smoothness. We then propose variations of the algorithms mentioned before that only require gradient and Hessian-vector product oracles, rather than gradient and Hessian oracle as well as a linear system solver. We also leverage the previous analysis in the case of inexact directions in order to provide computational complexity results for the second-order conjugate gradient type algorithms, which have the same order of $\epsilon$ as the results for the aforementioned exact methods. Our results demonstrate that our algorithms are more robust and the second-order conjugate gradient variations require less computational cost compared to existing methods.

### 2.1.2 Related Literature

We now discuss previous works related to our algorithm and analysis from three distinct aspects.

**Other Path Methods and Comparison of Results** As previous mentioned, for the LASSO and SVM problems, the exact solution path is piecewise linear and can be computed by the path following methods such as the least angle regression (LARS) algorithm proposed by [27, 39, 69]. Additional problems whose solution paths are piecewise linear are considered by [76], and [55] showed that the number of breakpoints in the solution path can be exponential in the number of data points. Generalized linear regression problems with $\ell_1$ regularization are considered in [86, 91] via LARS based methods. Another line of works focuses on computing approximate solution paths for specific problems, including the elastic net in [32], the SVM problem in [12, 35], matrix completion with nuclear norm regularization in [61], and other structural convex problems in [36, 53]. For problems with non-convex but coordinate decomposable regularization functions, a coordinate descent based algorithm was proposed by [60, 84]. Closest to our problem set up, [75] considered a general problem when $f(\cdot)$ and $\Omega(\cdot)$ have third-order derivatives and provided an algorithm which applied exact Newton steps on equally spaced grids starting from the optimal solution of the non-regularized problem. The lower bound complexity analysis when the approximate solution path is limited to piecewise constant function is considered by [35, 63].

**Related global complexity analysis of second-order methods** Newton-like methods are an important class of algorithms in optimization. Some notable lines of work include interior-point methods [65, 74] and applications in regression problems [46] as well as the Levenberg-Marquardt Method [62]. In practice, one often incorporates techniques to ensure global convergence, such as line searches [73] and trust-region techniques [23]. The global complexity analysis of Newton and higher-order methods with regularization has also received much recent interest, such as the work of [64, 66, 72] and the references therein. In

our paper, we make similar types of assumptions as in the global complexity analysis of regularized second and higher-order methods and we also prove global complexity results for the class of second-order algorithms considered herein.

**Related work on differential equations and optimization methods** Early works, including the work of [7, 8], analyzed inertial dynamical systems driven by gradient and Newton flow with application to optimization problems. Newton-like dynamic systems of monotone inclusions with connections to Levenberg-Marquardt and regularized Newton methosd were also analyzed by [1, 10, 21]. Due to the recent popularity of the accelerated gradient descent method in the machine learning and optimization communities, the limiting dynamics of different optimization methods and their discretization have thus received much-renewed interest in recent years; see [11, 77, 81, 85, 87] and the references therein.

### 2.1.3 Organization

This chapter is organized as follows. In Section 2.2, we derive the ordinary differential equation with an initial condition whose solution is the exact solution path (2.3) and provide the existence and the uniqueness of the solution of the differential equation. In Section 2.3, we leverage the ODE to develop a numerical algorithm to compute the approximate solution path of (2.2), and we derive the corresponding complexity analysis in Theorem 2.3.1. In Section 2.4, we propose a multi-stage method which is beneficial when the functions $f(\cdot)$ and $\Omega(\cdot)$ have higher-order smoothness, and we also provide its complexity analysis in Theorem 2.4.1. In Section 2.5, we discuss the cases with the presence of inexact oracles, and as a direct application we propose the second-order conjugate gradient variants of the aforementioned algorithms, which avoid exact Hessian evaluations and exact solutions of linear systems. Section 2.6 contains a detailed computational experiments of the proposed algorithms and grid search methods on both real and synthetic datasets.

### 2.1.4 Notation

For a positive integer $n$, let $[n] := \{1, \dots, n\}$. For a vector-valued function $y(t) : \mathbb{R} \to \mathbb{R}^p$ which can be written as $y(t) = (y_1(t), \dots, y_p(t))$, we say $y(\cdot)$ is differentiable if $y_i(\cdot)$ is differentiable for all $i = 1, \dots, p$ and let $\frac{\mathrm{d}y}{\mathrm{d}t}$ be the derivative of $y(t)$, namely $\frac{\mathrm{d}y}{\mathrm{d}t} = (\frac{\mathrm{d}y_1}{\mathrm{d}t}, \dots, \frac{\mathrm{d}y_p}{\mathrm{d}t})$. Let $\mathbf{1}_p$ and $\mathbf{1}_{p\times p}$ denote the $p$-dimensional all-ones vector and the $p \times p$ all-ones matrix respectively. Throughout the paper, we fix the norm $\|\cdot\|$ on $\mathbb{R}^p$ to be the $\ell_2$-norm, which is defined by $\|x\| := \|x\|_2 = \sqrt{x^T x}$. Also, in a slight abuse of notation, we use $\|\cdot\|$ to represent the operator norm, i.e., the induced $\ell_2$-norm $\|\cdot\|_2$ on $\mathbb{R}^{n\times p}$, which is defined by $\|A\| := \|A\|_{2,2} = \max_{\|x\|_2 \le 1} \|Ax\|_2$.

## 2.2 Ordinary Differential Equation Characterization of Solution Path

Let us describe a differential equations perspective on the solution path that will prove fruitful in developing efficient computational methods. First we introduce a re-parameterization in terms of an auxiliary variable $t \geq 0$ (thought of as "time"), whereby for a given $T > 0$ we introduce 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 $\frac{\mathrm{d}\lambda}{\mathrm{d}t} = \xi(\lambda(t))$ yields

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\left(\nabla^2 f(x(t)) + \lambda(t)\nabla^2\Omega(x(t))\right)^{-1} \xi(\lambda(t))\nabla\Omega(x(t)).$$

Then, apply the fact that $\nabla f(x(t)) + \lambda(t)\nabla\Omega(x(t)) = 0$ 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

$$\frac{\mathrm{d}\lambda}{\mathrm{d}t} = \xi(\lambda), \quad \frac{\mathrm{d}x}{\mathrm{d}t} = v(x, \lambda) := \left(\nabla^2 f(x) + \lambda\nabla^2\Omega(x)\right)^{-1} \frac{\xi(\lambda)}{\lambda}\nabla f(x), \tag{2.4}$$

for $t \in [0, T]$.

By considering specific choices of $\xi(\cdot)$ and $\Omega(\cdot)$, the system (2.4) generalizes some previously studied methodologies in parameteric optimization. First, consider the scenario with an equally-spaced discretization of the interval $[0, T]$, namely $t_k = k \cdot h$ for some fixed step-size $h > 0$. Thus, the sequence $\lambda_k := \lambda(t_k)$ is approximately given by $\lambda_{k+1} \approx \lambda_k + h \cdot \xi(\lambda_k)$. Intuitively, the choice of $\xi(\cdot)$ controls the dynamic of $\lambda(\cdot)$ and generalizes some previously considered sequences $\{\lambda_k\}$ for problem (2.2). For example, by letting $\xi(\lambda) \equiv 1$ we recover the arithmetic sequence in [75] and by letting $\xi(\lambda) \equiv -\lambda$ we recover the geometric sequence in [63]. In addition, consider the special case when $\Omega(x) = \frac{1}{2}\|x\|^2$. Then the dynamic for $x(t)$ in (2.4) is similar to the limiting dynamic for proximal Newton method (also known as the Levenberg-Marquardt regularization procedure [57] for convex optimization problems). The property of a similar dynamic of monotone inclusion is analyzed in [9, 10], which includes finding zero of the gradient of a convex function.

Before developing and presenting algorithms designed for computing the approximate solution path based on (2.4), we first verify that the system (2.4) has a unique trajectory. The proposition below states conditions on $f(\cdot)$ and $\Omega(\cdot)$ such that $v(\cdot,\cdot)$ defined in (2.4) is continuous in $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and is uniformly $L_v$-Lipschitz continuous with respect to $x$, namely,

$$\|v(x_1,\lambda) - v(x_2,\lambda)\| \le L_v\|x_1 - x_2\|, \quad \forall \lambda \in [\lambda_{\min}, \lambda_{\max}], x_1, x_2 \in \mathbb{R}^p.$$

This uniform Lipschitz property ensures that the above system has a unique trajectory, which therefore coincides with the solution path defined in (2.3).

**Proposition 2.2.1** (Theorem 5.3.1 of [34]). *If $\xi(\lambda)$ is Lipschitz continuous on $[\lambda_{\min}, \lambda_{\max}]$, $v(x,\lambda)$ is both continuous in $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and with respect to $x$ satisfies a uniform Lipschitz condition*

$$\|v(x_1,\lambda) - v(x_2,\lambda)\|_2 \le L_v\|x_1 - x_2\|_2, \quad \forall \lambda \in [\lambda_{\min}, \lambda_{\max}], x_1, x_2 \in \mathbb{R}^p,$$

*then (2.4) has a unique solution $(\lambda(t), x(t))$ for $t \in [0,T]$. Moreover, when $\nabla^2 f(\cdot)$, $\nabla^2\Omega(\cdot)$, $\nabla f(\cdot)$, $f(\cdot)$ are $L$-Lipschitz continuous, $f(\cdot)$ is $\mu$-strongly convex, $\Omega(\cdot)$ is $\sigma$-strongly convex, and $|\frac{\xi(\lambda)}{\lambda}| \le C$ for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, it holds that $v(\cdot,\cdot)$ defined in (2.4) is $L_v$-Lipschitz continuous with*

$$L_v = \frac{LC}{\mu + \lambda_{\min}\sigma} + \frac{L^2C(1+\lambda_{\max})}{(\mu+\lambda_{\min}\sigma)^2}.$$

## 2.3 Discretization and Complexity Analysis

In this section, we present algorithms for computing an approximate solution path based on discretizations of (2.4), along with the corresponding complexity analysis. The primary error metric that we consider is the 2-norm of the gradient across the entire interval $[\lambda_{\min}, \lambda_{\max}]$ as formally presented in Definition 2.3.1.

**Definition 2.3.1.** *An approximate solution path $\hat{x}(\cdot) : [\lambda_{\min}, \lambda_{\max}] \to \mathbb{R}^p$ to the parametric optimization problem (2.2) has accuracy $\epsilon \ge 0$ if $\|\nabla F_\lambda(\hat{x}(\lambda))\| \le \epsilon$ 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 $\epsilon$-accurate solution path $\hat{x}(\cdot)$ also has the optimality gap guarantee, which is

$$F_\lambda(\hat{x}(\lambda)) - F_\lambda^* \le \frac{\epsilon^2}{2(\mu+\lambda\sigma)} \le \frac{\epsilon^2}{2(\mu+\lambda_{\min}\sigma)},$$

for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. Algorithm 2.1 below presents a two-step "meta-algorithm" for computing an approximate solution path $\hat{x}(\cdot)$. Inspired by numerical methods to solve ordinary differential equations, we first design several schemes to iteratively update $(x_k, \lambda_k)$ by exploiting the dynamics in (2.4). We use the function $\psi(\cdot,\cdot) : \mathbb{R}^p \times [\lambda_{\min}, \lambda_{\max}] \to \mathbb{R}^p \times [\lambda_{\min}, \lambda_{\max}]$

**Algorithm 2.1:** Meta-algorithm for computing an approximate solution path $\hat{x}(\cdot)$

**input :** initial point $x_0 \in \mathbb{R}^p$, total number of iterations $K \geq 1$, update rule $\psi(\cdot, \cdot)$, and interpolation method $\mathcal{I}(\cdot)$

Initialize regularization parameter $\lambda_0 \leftarrow \lambda_{\max}$;

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

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

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

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 $\hat{x}(\cdot)$ over $[\lambda_{\min}, \lambda_{\max}]$.

We develop oracle complexity results for different update schemes and interpolation methods in terms of the number of gradient computations, Hessian computations, and linear systems solved required to compute an $\epsilon$-accurate approximate path. In this section, we will stick to a simple version of Algorithm 2.1 based on applying se mi-implicit Euler's method and linear interpolation to specify the update rule $\psi(\cdot, \cdot)$ and interpolation method $\mathcal{I}(\cdot)$. In particular, the implicit Euler's discretization of (2.4) is

$$\lambda_{k+1} = \lambda_k + h \cdot \xi(\lambda_k), \quad x_{k+1} = x_k + h \cdot v(x_k, \lambda_{k+1}), \tag{2.5}$$

and the linear interpolation $\mathcal{I}_{\text{linear}}(\cdot) : \{(x_k, \lambda_k)\}_{k=0}^K \to \hat{x}(\cdot)$ is defined by $\hat{x}(\lambda) := \alpha x_k + (1 - \alpha)x_{k+1}$ with $\alpha = \frac{\lambda - \lambda_{k+1}}{\lambda_k - \lambda_{k+1}}$ for all $\lambda \in [\lambda_{k+1}, \lambda_k]$ and $k \in \{0, \ldots, K-1\}$.

**The algorithm with the exponential decaying parameter sequence** Recall the update rule (2.5), we can see that the function $\xi(\cdot)$ (or equivalently, $\lambda(\cdot)$) is still to be determined. In practical cases, the value of $\lambda$ usually exponentially decreases from $\lambda_{\max}$ to $\lambda_{\min}$. This choice of penalty scale parameters $\{\lambda_k\}$ arises in the solution path for linear models, see [32], and the interior-point method, see [74]. Although our analysis holds for a broad class of $\lambda(\cdot)$, we first present the version with an exponential decaying parameter sequence, namely $\lambda(t) = e^{-t}\lambda(0)$. This specific version of Algorithm 2.1 is formally described in Algorithm 2.2.

Before going further into detailed analysis, we first state the computational guarantee of Algorithm 2.2. In our complexity analysis, we make the following smoothness assumptions on $f(\cdot)$ and $\Omega(\cdot)$.

**Assumption 2.3.1.** *In addition to $\mu$-strong convexity of $f(\cdot)$ and $\sigma$-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 $f^* := \min_x f(x) > -\infty$, 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 2.2:** Euler method for computing an approximate solution path $\hat{x}(\cdot)$

**input :** initial point $x_0 \in \mathbb{R}^p$, total number of iterations $K \geq 1$

Initialize regularization parameter $\lambda_0 \leftarrow \lambda_{\max}$, set step-size $h \leftarrow 1 - (\frac{\lambda_{\min}}{\lambda_{\max}})^{\frac{1}{K}}$;

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

  Update $\lambda_{k+1} \leftarrow (1-h)\lambda_k$;

  Update $x_{k+1} \leftarrow x_k - h\left(\nabla^2 f(x_k) + \lambda_{k+1}\nabla^2\Omega(x_k)\right)^{-1}\nabla f(x_k)$;

**output:** $\hat{x}(\cdot) \leftarrow \mathcal{I}_{\text{linear}}\left(\{(x_k, \lambda_k)\}_{k=1}^K\right)$ according to linear interpolation

Theorem 2.3.1 is our main result concerning the complexity of Algorithm 2.2 and demonstrates that in terms of the accuracy parameter $\epsilon$, Algorithm 2.2 requires $\mathcal{O}(1/\epsilon)$ iterations to compute an $\epsilon$-accurate solution path.

**Theorem 2.3.1.** *Suppose that Assumption 2.3.1 holds, let $\epsilon > 0$ be the desired accuracy, and suppose that the initial point $x_0$ satisfies $\|\nabla F_{\lambda_{\max}}(x_0)\| \leq \frac{\epsilon}{4}$. Let $T := \log(\lambda_{\max}/\lambda_{\min})$, let $\tau = \max\{\frac{1+\lambda_{\min}}{\mu+\lambda_{\min}\sigma}, \frac{1+\lambda_{\max}}{\mu+\lambda_{\max}\sigma}\}$, and let*

$$K_{\mathrm{E}} := \left\lceil \max\left\{2T, \frac{\sqrt{LG}\tau T}{\sqrt{3}}, \frac{4(f(x_0)-f^*)\tau LT}{\epsilon}, \frac{2\sqrt{L}(\tau G+1)T}{\sqrt{\epsilon}}\right\}\right\rceil. \tag{2.6}$$

*If the total number of iterations $K$ satisfies $K \geq K_{\mathrm{E}}$, then Algorithm 2.2 returns an $\epsilon$-accurate solution path.*

**Remark 2.3.1.** *Grid search type methods for computing approximate solution paths are proposed in [35, 63], and we will follow the analysis in the latter one, which considers more general cases. In order to guarantee the function value gap $h(x) - h^* \leq \epsilon'$, it will require the number of grid points $K = \frac{\sqrt{\tau}GT}{\sqrt{\epsilon'}}$. For $L$-smooth function $h(\cdot)$, $h(x) - h^* \leq \frac{\epsilon^2}{2L}$ implies $\|h(x)\| \leq \epsilon$, which is the goal in our paper. Therefore, we need to set $\epsilon' = \frac{\epsilon^2}{2L}$, and hence we have the number of grid points is $K = \frac{\sqrt{\tau L}GT}{\epsilon}$ when the desired accuracy of the inner problem is set to $\epsilon_c = \frac{\epsilon'}{2}$. Using the exact Newton's method and the last grid point as a warm-start to solve the inner problem, [45] implies that the inner complexity is $(\tau L)^2 \log 2$, and therefore, the total complexity is $\frac{(\tau L)^{5/2}GT\log 2}{\epsilon}$.*

From the results in Theorem 2.3.1 and Remark 2.3.1, we can see that both our results and the grid search method have the complexity of order $\mathcal{O}(\frac{1}{\epsilon})$. In most practical cases, $f(x_0) - f^*$ is smaller than $(\tau L)^{3/2}G\log 2$, which implies the constant in (2.6) is better. In the later part of this paper, we will propose algorithms utilizing the smoothness of $f(\cdot)$ and $\Omega(\cdot)$ and achieve better complexity results.

### 2.3.1 Semi-Implicit Euler Update Scheme

Herein we provide the computational guarantee of the semi-implicit Euler update scheme defined in (2.5), wherein the main object we concern is the accuracy at each grid point. Let $r_k := \|\nabla F_{\lambda_k}(x_k)\|$ denote the accuracy at $(x_k, \lambda_k)$. We consider a broad family of $\xi(\cdot)$ in the following analysis, although in Algorithm 2.2 and the corresponding complexity analysis in Theorem 2.3.1, we only consider the special scenario that $\lambda(t) = e^{-t}\lambda(0)$. We first present the computation guarantees of Taylor expansion approximations given the Lipschitz continuity of Hessian in Lemma 2.3.1.

**Lemma 2.3.1** (Lemma 1 in [66])**.** *Suppose $\phi(\cdot) : S \to \mathbb{R}$ has L-Lipschitz Hessian for some convex set $S \subseteq \mathbb{R}^d$, then the following inequalities hold for all $x, y \in S$:*

*(i)* $\|\nabla\phi(y) - \nabla\phi(x) - \nabla^2\phi(x)(y-x)\| \le \frac{1}{2}L\|y-x\|^2$.

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

Based on the results in Lemma 2.3.1, we provide the local analysis of $r_k$, that is, how the norm of the gradient at each grid point accumulates. Lemma 2.3.2 provides an upper bound on $\|r_{k+1}\|$ based on $\|r_k\|$, which represents the accuracy at the previous iteration.

**Lemma 2.3.2.** *Suppose Assumption 2.3.1 holds, discretization* (2.5) *has the following guarantee for all $k \ge 0$:*

$$r_{k+1} \le \frac{\lambda_{k+1}}{\lambda_k} \cdot r_k + h^2 \cdot \frac{L(1+\lambda_{k+1})}{2} \|v(x_k, \lambda_{k+1})\|^2 .$$

*Proof.* It holds that

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

where the first inequality is true because of the triangle inequality, and in the second inequality, for the first two terms, we apply Item i in Lemma 2.3.1, and for the third terms, they are equal to each other. □

Lemma 2.3.2 provides the first technical result of semi-implicit Euler's update scheme. When Assumption 2.3.1 holds (2.5) has the computation guarantee $\frac{r_{j+1}}{\lambda_{j+1}} \le \frac{r_j}{\lambda_j} + \mathcal{O}(h^2)$. After

telescoping the inequalities for $j = 0, 1, \ldots, k-1$ for $k \leq K = T/h$ we have $r_k \sim \mathcal{O}(h)$. Hence, one would expect a uniform $\mathcal{O}(h)$ bound on all accuracy $r_k$ at points $x_k$ for all $k = 0, \ldots, K$. We formalize the idea in the following lemma.

**Lemma 2.3.3.** *Suppose Assumption 2.3.1 holds, $\lambda_{k+1} \geq \lambda_{\min}$, $\xi(\lambda_j) < 0$ for all $j \leq k$, and step-size $h$ satisfies the following condition for all $0 \leq j < k$:*

$$h \leq \min\left\{ \frac{\lambda_j}{-2\xi(\lambda_j)}, \sqrt{\frac{3\lambda_j(\mu + \lambda_{j+1}\sigma)^2}{-\xi(\lambda_j)LG}} \right\}. \tag{2.7}$$

*Then the sequence $\{(x_k, \lambda_k)\}_{k=0}^K$ generated by update scheme (2.5) satisfies $f(x_{k+1}) \leq f(x_k)$, and the corresponding accuracy $\{r_k\}_{j=0}^K$ has the following guarantee:*

$$\frac{r_k}{\lambda_k} \leq \frac{r_0}{\lambda_0} + 2hL(f(x_0) - f(x_k)) \cdot \max_{j \in [k-1]} \left\{ \frac{-(1+\lambda_{j+1})\xi(\lambda_j)}{\lambda_j \lambda_{j+1}(\mu + \lambda_{j+1}\sigma)} \right\}. \tag{2.8}$$

*Proof.* We will use induction to show that $f(x_{j+1}) \leq f(x_j)$, which implies that $x_k \in S_{x_0}$ for all $k \in \{0, \ldots, K\}$. First, suppose $x_j \in S_{x_0}$ for some $j \in \{0, \ldots, K-1\}$. Let $d_j$ denote $v(x_j, \lambda_{j+1})$ and it holds that $\|d_j\| \leq \frac{G}{\mu + \lambda_{j+1}\sigma}$. Since $h + \frac{\lambda_k}{2\xi(\lambda_k)} \leq 0$ and $h^2 \leq \frac{3\lambda_j(\mu+\lambda_{j+1}\sigma)^2}{-\xi(\lambda_j)LG} \leq \frac{3\lambda_j(\mu+\lambda_{j+1}\sigma)}{-\xi(\lambda_k)L\|d_k\|}$, it holds that

$$\begin{aligned} f(x_{j+1}) &\leq f(x_j) + h \cdot \nabla f(x_j)^T d_j + h^2 \cdot \frac{1}{2} d_j^T \nabla^2 f(x_j) d_j + h^3 \cdot \frac{1}{6} L \|d_j\|^3 \\ &\leq f(x_j) + \frac{h}{4} \cdot \frac{\lambda_j}{\xi(\lambda_j)} \cdot (\mu + \lambda_{j+1}\sigma) \|d_j\|^2 . \end{aligned}$$

Since $\xi(\lambda_j) < 0$, it holds that $f(x_{j+1}) \leq f(x_j)$ and therefore implies that $x_{j+1} \in S_{x_0}$. Then by induction we conclude that $x_j \in S_{x_0}$ for all $j \leq k$. Also, combining the result in Lemma 2.3.2, for all $j \leq k$, it holds that

$$\begin{aligned} \frac{r_{j+1}}{\lambda_{j+1}} &\leq \frac{r_j}{\lambda_j} + h^2 \cdot \frac{L(1+\lambda_{j+1})}{2\lambda_{j+1}} \|d_j\|^2 \\ &\leq \frac{r_j}{\lambda_j} + h \cdot \frac{L(1+\lambda_{j+1})}{2\lambda_{j+1}} \cdot \frac{4(f(x_j) - f(x_{j+1}))}{\frac{\lambda_j}{-\xi(\lambda_j)} \cdot (\mu + \lambda_{j+1}\sigma)}. \end{aligned}$$

By taking the summation over $j$ from 0 to $k$, we obtain (2.8). □

Lemma 2.3.3 provides the computation guarantees of Algorithm 2.2 with any decreasing sequence of $\{\lambda_k\}$. While in the upper bound on the accuracy at point $x_k$, it still involves a constant related to sequence $\{\lambda_k\}$. We then consider a family of the sequence $\{\lambda_k\}$ and then derive a simpler upper bound.

**Proposition 2.3.1.** *Suppose Assumption 2.3.1 holds,* $\lambda_{k+1} \geq \lambda_{\min}$, $-\lambda \leq \xi(\lambda) < 0$ *for all* $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, *and step-size* $h$ *satisfies*

$$h \leq \min\left\{\frac{1}{2}, \sqrt{\frac{3}{\tau^2 LG}}\right\}. \tag{2.9}$$

*Then under Assumption 2.3.1, discretization* (2.5), *it holds that*

$$r_{k+1} \leq r_0 + 2h\tau L(f(x_0) - f(x_{k+1})). \tag{2.10}$$

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

$$\frac{\lambda_j}{-2\xi(\lambda_j)} \geq \frac{1}{2} \quad \text{and} \quad \sqrt{\frac{3\lambda_j(\mu+\lambda_{j+1}\sigma)^2}{-\xi(\lambda_j)LG}} \geq \sqrt{\frac{3(\mu+\lambda_{\min}\sigma)^2}{LG}},$$

and hence condition (2.9) implies condition (2.7). Also, since $\frac{1+\lambda}{\mu+\lambda\sigma}$ is monotone in $[\lambda_{\min}, \lambda_{\max}]$, it holds that $\tau = \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \frac{1+\lambda}{\mu+\lambda\sigma}$. Therefore, we have

$$\max_{j\in[k-1]} \left\{\frac{-(1+\lambda_{j+1})\xi(\lambda_j)}{\lambda_j\lambda_{j+1}(\mu+\lambda_{j+1}\sigma)}\right\} \leq \frac{1+\lambda_k}{\lambda_k(\mu+\lambda_k\sigma)} \leq \frac{\tau}{\lambda_k}.$$

Apply the above inequality to (2.8) we obtain (2.10). □

Proposition 2.3.1 provides a uniform upper bound on accuracy of all near-optimal solutions $x_k$. When $\frac{\xi(\lambda)}{\lambda} \in [-1, 0)$ for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, Algorithm 2.2 generates a solution sequence $\{x_k\}$ such that $\|\nabla F_{\lambda_k}(x_k)\| \sim \mathcal{O}(h) + r_0$.

### 2.3.2 Linear Interpolation

In the last section, we provide a general accuracy analysis at all near-optimal solutions $x_k$. In this section, we analyze the second procedure in Algorithm 2.2, i.e., linear interpolation. Then the next step is to construct the entire path $\hat{x}(\lambda) : [\lambda_{\min}, \lambda_{\max}] \to \mathbb{R}^P$ based on these near-optimal solutions. First we recall the definition of linear interpolation for $x(\cdot), \lambda(\cdot)$:

$$\hat{\lambda}(t) := \alpha\lambda_k + (1-\alpha)\lambda_{k+1}, \quad \hat{x}(t) := \alpha x_k + (1-\alpha)x_{k+1}, \tag{2.11}$$

where $\alpha := \frac{t_{k+1}-t}{h}$ and $t \in [t_k, t_{k+1}]$ for all $k = 0, \ldots, K-1$. That is, given an arbitrary $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, we first select $t$ such that $\hat{\lambda}(t) = \lambda$, then we output $\hat{x} := \hat{x}(t)$ as a near-optimal solution to problem $P(\lambda)$. The following lemma provides an upper bound of $\|\nabla F_\lambda(\hat{x}(\lambda))\|$ for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$.

**Theorem 2.3.2.** *Suppose Assumption 2.3.1 holds. Let* $r_{\max} = \max_{0\leq k\leq K} r_k$. *Then, linear interpolation* $\hat{x}(\cdot), \hat{\lambda}(\cdot)$ *of sequence* $\{(x_k, \lambda_k)\}_{k=0}^K$ *generated by* (2.11) *has the following computational guarantee for all* $t \in [t_0, t_K]$*:*

$$\left\|\nabla F_{\hat{\lambda}(t)}(\hat{x}(t))\right\| \leq r_{\max} + \frac{L}{8} \cdot \max_{k\in[K-1]} \left\{(1+\lambda_k)\|x_{k+1}-x_k\|^2 + 2h|\xi(\lambda_k)|\|x_{k+1}-x_k\|\right\}.$$

*Proof.* First suppose $t \in [t_k, t_{k+1}]$. For simplicity, we define $x := \hat{x}(t)$, $\lambda := \hat{\lambda}(t)$, $\delta_1 := \nabla f(x_k) + \lambda_k \nabla\Omega(x_k)$, $\delta_2 := \nabla f(x_{k+1}) + \lambda_{k+1}\nabla\Omega(x_{k+1})$. By triangle inequality and Item i in Lemma 2.3.1, it holds that

$$\begin{aligned}\|\alpha\nabla f(x_k) + (1-\alpha)\nabla f(x_{k+1}) - \nabla f(x)\| &\leq \frac{\alpha(1-\alpha)L}{2}\|x_k - x_{k+1}\|^2 \\ &\leq \frac{L}{8}\|x_k - x_{k+1}\|^2 .\end{aligned}$$

Also, by applying similar trick on $\nabla\Omega(\cdot)$, it holds that Then,

$$\begin{aligned}&\|\lambda\nabla\Omega(x) - \alpha\lambda_k\nabla\Omega(x_k) - (1-\alpha)\lambda_{k+1}\nabla\Omega(x_{k+1})\| \\ \leq &\|\lambda(\alpha\nabla\Omega(x_k) + (1-\alpha)\nabla\Omega(x_{k+1}) - \nabla\Omega(x))\| \\ &+ \|\alpha(\lambda-\lambda_k)\nabla\Omega(x_k) + (1-\alpha)(\lambda-\lambda_{k+1})\nabla\Omega(x_{k+1})\| \\ \leq &\frac{\lambda L}{2}\alpha(1-\alpha)\|x_{k+1}-x_k\|^2 + \|\alpha(1-\alpha)(\lambda_{k+1}-\lambda_k)(\nabla\Omega(x_{k+1}) - \nabla\Omega(x_k))\| \\ \leq &\frac{\lambda L}{8}\|x_{k+1}-x_k\|^2 + \frac{h\,|\xi(\lambda_k)|\,L}{4}\|x_{k+1}-x_k\| .\end{aligned}$$

Combine the above two inequality and apply triangle inequality, we have

$$\begin{aligned}&\|\nabla f(x) + \lambda\nabla\Omega(x) - \alpha\delta_1 - (1-\alpha)\delta_2\| \\ \leq &\frac{L}{8}\|x_{k+1}-x_k\|^2 + \frac{\lambda L}{8}\|x_{k+1}-x_k\|^2 + \frac{|\xi(\lambda_k)|\,Lh}{4}\|x_{k+1}-x_k\| .\end{aligned}$$

□

Theorem 2.3.2 provides the computational guarantee of linear interpolation of the sequence $\{(x_k, \lambda_k\}_{k=0}^K$ generated by update scheme (2.5). Observed that $\|x_{k+1} - x_k\|$ is of the order $\mathcal{O}(h)$, we have that additional error incurred by linear interpolation is of the order $\mathcal{O}(h^2)$. Together with the $\mathcal{O}(h)$ accuracy from the update scheme (2.5), we are able to provide a computational guarantee on the accuracy of the approximate path generated by Algorithm 2.2. In the following part we will provide a uniform bound on $\|\nabla F_{\hat{\lambda}(t)}(\hat{x}(t))\|$ for a family of $\lambda(t)$ and discretization.

### 2.3.3 Computational Guarantee for the Exponential Decaying Parameter Sequence

Under the $\lambda(t) = \lambda_{\max}\cdot\exp(-t)$ scenario, we have $\xi(\lambda) = -\lambda$ and $\xi(\cdot)$ satisfies the assumption in Proposition 2.3.1. Therefore, we extend the result in Theorem 2.3.2 and provide an explicit uniform bound of the path accuracy.

**Proposition 2.3.2.** *Suppose Assumption 2.3.1 holds, and the step-size $h$ satisfies the condition in* (2.9)*. Let $\hat{x}(\cdot) : [0,T] \to \mathbb{R}^p$ and $\hat{\lambda}(\cdot) : [0,T] \to [\lambda_{\min}, \lambda_{\max}]$ denote the approximate*

*solution path generated by Algorithm 2.2. Let $f^*$ denote the minimum value of $f(\cdot)$. Then, we have the following computational guarantee for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$:*

$$\|\nabla F_\lambda(\hat{x}(\lambda))\| \leq \|\nabla F_{\lambda_{\max}}(x_0)\| + 2h\tau L(f(x_0) - f^*) + \frac{h^2 L}{8} \cdot (\tau G + 1)^2. \tag{2.12}$$

*Proof.* First we extend the result in Proposition 2.3.1, and it holds that

$$r_{\max} \leq \max_{k\in[K]} r_k \leq r_0 + 2hL(f(x_0) - f^*) \cdot \tau.$$

Also, for the result in Theorem 2.3.2, we further have

$$\begin{aligned}
&\frac{L}{8} \cdot \max_{k\in[K-1]} \left\{(1+\lambda_k)\|x_{k+1} - x_k\|^2 + 2h|\xi(\lambda_k)|\|x_{k+1} - x_k\|\right\} \\
\leq &\frac{h^2 L}{8} \cdot \max_{k\in[K-1]} \left\{ \frac{(1+\lambda_k)G^2}{(\mu + \lambda_k\sigma)^2} + \frac{2\lambda_k G}{\mu + \lambda_k \sigma} \right\} \leq \frac{h^2 L}{8} \cdot (\tau G + 1)^2.
\end{aligned}$$

Combine the above two inequalities and Proposition 2.3.1 and theorem 2.3.2, we obtain (2.12). □

In Algorithm 2.2, the sequence $\{\lambda_k\}_{k=0}^K$ is given by $\lambda_{k+1} = (1-h)\lambda_k$, and it implies that $\lambda_{\min} = (1-h)^K \lambda_{\max}$. Hence, we have $h = 1 - (\frac{\lambda_{\min}}{\lambda_{\max}})^{1/K}$. Apply the fact to Proposition 2.3.2, we arrive at the complexity analysis with respect to the number of iteration. We formalize the complexity analysis in the following proof of Theorem 2.3.1, which appears at the beginning of this section.

*Proof of Theorem 2.3.1.* The conditions that $K \geq \max\left\{2T, \frac{\sqrt{LG}\tau T}{\sqrt{3}}\right\}$ and $h = 1 - \left(\frac{\lambda_{\min}}{\lambda_{\max}}\right)^{1/K} \leq \frac{T}{K}$ guarantee that step-size $h$ satisfies (2.9). Also $K \geq \frac{\tau LT(f(x_0)-f^*)}{\epsilon}$ and $K \geq \frac{(\tau G+1)\sqrt{LT}}{\sqrt{\epsilon}}$ guarantees that $2h\tau L(f(x_0) - f^*) \leq \frac{\epsilon}{2}$ and $\frac{h^2 L}{8} \cdot (\tau G + 1)^2 \leq \frac{\epsilon}{4}$. Hence Algorithm 2.2 guarantees a $\epsilon$-accurate solution path. □

Recall that in the assumption of Theorem 2.3.1, it requires a good initialization $x_0$ satisfying $\|\nabla F_{\lambda_{\max}}(x_0)\| \leq \frac{\epsilon}{2}$. In practical cases, we can either implement a specific convex optimization algorithm to get an $x_0$ satisfying the initial condition or use the initialization suggested in the following lemma. Here we suggest one choice of initialization with computational guarantee when the function $\Omega(\cdot)$ is structured, i.e., the minimizer of $\Omega(\cdot)$ is easy to calculate.

**Lemma 2.3.4.** *Suppose Assumption 2.3.1 holds. Let initialization $x_0$ be*

$$x_0 := x_\Omega - \left(\nabla^2 f(x_\Omega) + \lambda_{\max}\nabla^2\Omega(x_\Omega)\right)^{-1} \nabla f(x_\Omega),$$

*where $x_\Omega := \arg\min_{x\in\mathbb{R}^p} \Omega(x)$, then it holds that*

$$\|\nabla f(x_0) + \lambda_{\max}\nabla\Omega(x_0)\| \leq \frac{L(1+\lambda_{\max})\|\nabla f(x_\Omega)\|^2}{2(\mu + \lambda_{\max}\sigma)^2}.$$

*Proof.* Define $d = -\left(\nabla^2 f(x_\Omega) + \lambda_{\max}\nabla^2\Omega(x_\Omega)\right)^{-1}\nabla f(x_\Omega)$, it holds that

$$\begin{aligned}\|\nabla f(x_0) + \lambda_{\max}\nabla\Omega(x_0)\| &\le \left\|\nabla f(x_\Omega) + \nabla^2 f(x_\Omega)d + \lambda_{\max}\left(\nabla\Omega(x_\Omega) + \nabla^2\Omega(x_\Omega)d\right)\right\| \\ &\quad + \frac{(1+\lambda_{\max})L}{2}\|d\|^2 \le \frac{L(1+\lambda_{\max})\|\nabla f(x_\Omega)\|^2}{2(\mu+\lambda_{\max}\sigma)^2},\end{aligned}$$

where the first inequality follows Lemma 2.3.1, and the second inequality holds since $\nabla\Omega(x_\Omega) = 0$, $\left(\nabla^2 f(x_\Omega) + \lambda_{\max}\nabla^2\Omega(x_\Omega)\right)d + \nabla f(x_\Omega) = 0$, and $\|d\| \le \frac{\|\nabla f(x_\Omega)\|}{\mu+\lambda_{\max}\sigma}$. □

Notice that the value of $L$ and $\|\nabla f(x_\Omega)\|$ are independent of $\lambda_{\max}$. Hence, when $\lambda_{\max}$ is sufficiently large, we have $r_0 \le \frac{\epsilon}{4}$. Also, since $x_\Omega$ is the optimal solution when $\lambda = +\infty$, the initialization can be regarded as an updating step of (2.5) from $x_\Omega$.

## 2.4 Multi-Stage Discretization

In the analysis of the previous section, Algorithm 2.2 guarantees an $\epsilon$-accurate solution path within $\mathcal{O}(\epsilon^{-1})$ calls to the gradient, Hessian oracle, and linear equations solver. One advantage of the main result proposed in Theorem 2.3.1 is that only the smooth Hessian of $f(\cdot)$ and $\Omega(\cdot)$ is required and no assumption of $\epsilon$ is required. When $f(\cdot)$ and $\Omega(\cdot)$ have better properties and $\epsilon$ is relatively small, one would like an algorithm which utilizes these properties and requires fewer calls to oracle with respect to the order of $\epsilon$. Motivated by the multi-stage numerical methods for solving differential equations, we design several update schemes to achieve higher-order accuracy. Specially, in this section we still consider the exponentially decaying parameter, that is, $\lambda(t) = e^{-t}\lambda(0)$.

### 2.4.1 Trapezoid Method

In this section, we propose and analyze the *trapezoid method*, whose formal description is given in Algorithm 2.3. The trapezoid method is beneficial when the function $f(\cdot)$ and $\Omega(\cdot)$ have Lipschitz continuous third-order derivatives and it achieves a higher-order accuracy than the implicit Euler method. The accuracy of the output path by Algorithm 2.3 has the order $\mathcal{O}(h^2)$ where $h$ is the step-size, or equivalently, $\mathcal{O}(K^{-2})$ where $K$ is the number of iterations. Moreover, we want to mention that the algorithm does not require the oracle to higher-order derivatives, but still only requires gradient and Hessian oracle as in the Euler method.

**Algorithm 2.3:** Trapezoid method for solution path

**input :** Initial point $x_0 \in \mathbb{R}^p$, total number of iterations $K \geq 1$

Initialize parameter $\lambda_0 \leftarrow \lambda_{\max}$, set step-size $h \leftarrow 1 - \sqrt{2(\frac{\lambda_{\min}}{\lambda_{\max}})^{\frac{1}{K}} - 1}$;

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

  Update $d_{k,1} \leftarrow v(x_k, \lambda_k)$;

  Update $d_{k,2} \leftarrow v(x_k + h \cdot d_{k,1}, (1 - h + h^2)\lambda_k)$;

  Update $x_{k+1} \leftarrow x_k + h \cdot \frac{d_{k,1}+d_{k,2}}{2}$;

  Update $\lambda_{k+1} \leftarrow (1 - h + \frac{h^2}{2})\lambda_k$;

**output:** $\hat{x}(\cdot) \leftarrow \mathcal{I}_{\text{linear}}\left(\{(x_k, \lambda_k)\}_{k=0}^{K}\right)$ according to linear interpolation

We first state the main technical assumption and computational guarantees of Algorithm 2.3.

**Assumption 2.4.1.** *In addition to Assumption 2.3.1, we assume the third-order directional derivative of $f(\cdot)$ and $\Omega(\cdot)$ are L-Lipschitz continuous and $\sigma \geq 1$.*

**Theorem 2.4.1.** *Suppose Assumption 2.4.1 holds, let $\epsilon > 0$ be desired accuracy, let $\tilde{\mu} := \mu + \lambda_{\min}\sigma$, suppose that the initial point $x_0$ satisfies $\|\nabla F_{\lambda_{\min}}(x_0)\| \leq \frac{\epsilon}{2} \leq \tilde{\mu}$, let $T := 1.1\log(\lambda_{\max}/\lambda_{\min})$, and let*

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

*If the total number of iterations $K$ satisfies $K \geq K_{\text{tr}}$, then Algorithm 2.3 returns an $\epsilon$-accurate solution path.*

The result in Theorem 2.4.1 shows that we improve the total complexity to $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$, which is better the $\mathcal{O}(\frac{1}{\epsilon})$ complexity of the Euler method and the best known results in grid search method (see Theorem 2.3.1 and remark 2.3.1). Similar as previous complexity analysis of the semi-implicit Euler method, the analysis of the trapezoid method consists of two part: We first present the computation guarantee of trapezoid update scheme, which is defined as

$$(x_{k+1}, \lambda_{k+1}) \leftarrow T(x_k, \lambda_k) := \left(x_k + h \cdot \tfrac{d_{k,1}+d_{k,2}}{2}, (1 - h + \tfrac{h^2}{2})\lambda_k\right), \tag{2.13}$$

where $d_{k,1} = v(x_k, \lambda_k)$, and $d_{k,2} = v(x_k + h \cdot d_{k,1}, (1 - h + h^2)\lambda_k)$.

**Lemma 2.4.1.** *Suppose Assumption 2.4.1 holds, $r_k = \|\nabla F_{\lambda_k}(x_k)\| \leq \tilde{\mu}$, and the next iterate $(x_{k+1}, \lambda_{k+1})$ is given by $(x_{k+1}, \lambda_{k+1}) = T(x_k, \lambda_k)$ defined in (2.13). Then, it holds that $(1 + \lambda_k)\|x_k - x_{k+1}\| \leq 3h(1+G)$, and*

$$r_{k+1} := \left\|\nabla F_{\lambda_{k+1}}(x_k)\right\| \leq \frac{\lambda_{k+1}}{\lambda_k} \cdot r_k + h^3 \cdot 3L(1+G)^3 + h^4 \cdot 2L^3\tau^2(1+G)^4. \tag{2.14}$$

We leave the proof of Lemma 2.4.1 in the appendix due to its length and complicacy. In the proof, we mainly work with the directional derivatives and the accuracy of Taylor expansion in Lemma 2.3.1. The result in Lemma 2.4.1 shows that trapezoid update scheme in (2.13) guarantees an $\mathcal{O}(h^3)$ local accumulation. Moreover, we can derive an $\mathcal{O}(h^2)$ uniform upper bound on accuracy of all near-optimal solutions $\{x_k\}$. For all other $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, we implement linear interpolation to approximate the corresponding near-optimal solution. We then provide the formal proof of Theorem 2.4.1.

*Proof of Theorem 2.4.1.* Since $h - \frac{h^2}{2} \leq \frac{1}{K}\log(\frac{\lambda_{\max}}{\lambda_{\min}})$, it holds that

$$h \leq \min\left\{0.1, \frac{\tilde{\mu}}{8L(1+G)}, \frac{\epsilon^{1/2}}{3L^{1/2}(1+G)^{3/2}}, \frac{\epsilon^{1/3}}{3\tau^{2/3}L(1+G)^{4/3}}\right\}.$$

Then we show that $r_k := \|\nabla F_{\lambda_k}(x_k)\| \leq \frac{\epsilon}{2}$ for all $k$ by induction. Suppose $r_k \leq \frac{\epsilon}{2}$, then by Lemma 2.4.1, it holds that

$$r_{k+1} := \left\|\nabla F_{\lambda_{k+1}}(x_k)\right\| \leq \frac{\lambda_{k+1}}{\lambda_k} \cdot r_k + h^3 \cdot 3L(1+G)^3 + h^4 \cdot 2L^3\tau^2(1+G)^4 \leq \frac{\epsilon}{2}.$$

Therefore, $r_k \leq \frac{\epsilon}{2}$ for all $k \in \{0, \ldots, K\}$. Suppose $\lambda \in [\lambda_{k+1}, \lambda_k]$, and hence $\hat{x}(\lambda) = \alpha x_k + (1-\alpha)x_{k+1}$ where $\alpha = \frac{\lambda - \lambda_{k+1}}{\lambda_k - \lambda_{k+1}}$. By applying results in Theorem 2.3.2, we have

$$\begin{aligned}\|f(\hat{x}(\lambda)) + \lambda\hat{x}(\lambda)\| &\leq \frac{\epsilon}{2} + \frac{L}{8}\max_{k\in[K-1]}\left\{(1+\lambda_k)\|x_{k+1}-x_k\|^2 + 2h|\xi(\lambda_k)|\|x_{k+1}-x_k\|\right\} \\ &\leq \frac{\epsilon}{2} + \frac{L}{8}\cdot\left(9h^2(1+G)^2 + 6h^2(1+G)\right) \leq \epsilon.\end{aligned}$$

□

### 2.4.2 Runge-Kutta Method

We describe the *Runge-Kutta* method for solution path in Algorithm 2.4, where at each iteration it requires 4 calls to Hessian oracle and linear equation solver. The interpolation function *CubicSpline* in the last step in Algorithm 2.4 is the *cubic spline interpolation*, which we will discuss later. The update rules of $(y_{k,\cdot}, \psi_{k,\cdot})$ and $(x_{k+1}, \lambda_{k+1})$ at each iteration come from the *classical Runge-Kutta method*. By calculating the Taylor series coefficients of the Runge-Kutta solution we can obtain the order of any arbitrary Runge-Kutta method. The classical Runge-Kutta method, as a special case, guarantees an $\mathcal{O}(h^5)$ local residual and hence an $\mathcal{O}(h^4)$ global residual. We omit the proof here and for detailed analysis of the Runge-Kutta method we refer readers to [22].

**Algorithm 2.4:** Runge-Kutta method for solution path

**input :** Initial point $x_0 \in \mathbb{R}^p$, total number of iterations $K \geq 1$

Initialize regularization parameter $\lambda_0 \leftarrow \lambda_{\max}$;

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

  Update $(y_{k,1}, \psi_{k,1}) \leftarrow (u(x_k, \lambda_k), -\lambda_k)$;

  Update $(y_{k,2}, \psi_{k,2}) \leftarrow (u(x_k + \frac{h}{2} \cdot y_{k,1}, \lambda_k + \frac{h}{2} \cdot \psi_{k,1}), -\lambda_k - \frac{h}{2} \cdot \psi_{k,1})$;

  Update $(y_{k,3}, \psi_{k,3}) \leftarrow (u(x_k + \frac{h}{2} \cdot y_{k,2}, \lambda_k + \frac{h}{2} \cdot \psi_{k,2}), -\lambda_k - \frac{h}{2} \cdot \psi_{k,2})$;

  Update $(y_{k,4}, \psi_{k,4}) \leftarrow (u(x_k + h \cdot y_{k,3}, \lambda_k + h \cdot \psi_{k,3}), -\lambda_k - h \cdot \psi_{k,3})$;

  Update $x_{k+1} \leftarrow x_k + \frac{h}{6} \cdot (y_{k,1} + 2y_{k,2} + 2y_{k,3} + y_{k,4})$;

  Update $\lambda_{k+1} \leftarrow \lambda_k + \frac{h}{6} \cdot (\psi_{k,1} + 2\psi_{k,2} + 2\psi_{k,3} + \psi_{k,4})$;

**output:** $\hat{x}(\cdot) \leftarrow \mathcal{I}_{\text{cubic}}\left(\{(x_k, \lambda_k)\}_{k=0}^K\right)$ according to cubic interpolation

In Algorithm 2.2 and Algorithm 2.3, we implement linear interpolation to recover the entire solution path. Theorem 2.3.2 guarantees an $\mathcal{O}(h^2)$ residual of linear interpolation. To increase the accuracy, we implement the *cubic spline interpolation*, which is a piece-wise third-order polynomial approximation. Suppose the problem is to find $\hat{x}(t_i) = x_i$ where $x_i = x(t_i)$ for $t_i = i \cdot h, i = 0, \ldots, K$. The cubic spline $\hat{x}(t)$ satisfies $\hat{x}(t) = p_i(t)$ for $x \in [t_{i-1}, t_i]$, where $\{p_i(\cdot)\}_{i=1}^K$ are $K$ cubic functions satisfying

$$\begin{aligned}
&p_i(t_i) = x_i, && p_i(t_{i+1}) = x_{i+1}, && i = 1, \ldots, K,\\
&p_i'(t_i) = p_{i+1}'(t_i), && p_i''(t_i) = p_{i+1}''(t_i), && i = 1, \ldots, K-1,\\
&p_1''(t_0) = 0, && p_K''(t_K) = 0.
\end{aligned}$$

One finds there are $4K(= 2K + 2(K-1) + 2)$ equations and $4K$ coefficients to be determined, which are 4 coefficients in each polynomial. The cubic spline interpolation guarantees $\|\hat{x}(t) - x(t)\| \leq O(h^4)$ for all $t \in [t_0, t_K]$. For other properties and detailed discussion of cubic spline interpolation we refer readers to [34].

## 2.5 Analysis with Inexact Linear Equations Solutions and Second-Order Conjugate Gradient Variants

In this section, we present the complexity analysis of the aforementioned methods with the presence of inexact oracle to gradient and Hessian and/or inexact linear equations solutions as well as variants applying the second-order conjugate gradient (SOCG) type methods. We first consider the case when gradient and Hessian oracle are inexact and/or linear equations solver yields approximate solutions.

### 2.5.1 Analysis with Inexact Oracle and/or Approximate Solver

At each iteration of the Euler, trapezoid, and Runge-Kutta method, one essential subroutine is to compute the directions $d_{k,i}$. For example, in the Euler method, $d_k$ is given

by the formula $d_k = v(x_k, \lambda_k) = -(\nabla^2 f(x_k) + \lambda_{k+1}\nabla^2\Omega(x_k))^{-1}\nabla f(x_k)$. In most large-scale problems, the computation of the Hessian matrix and solving linear equations exactly could be the computational bottleneck. Therefore, we consider the case with the presence of numerical error, which may be induced by inexact gradient and Hessian oracle, or by linear equations solver. Nevertheless, we tackle the two types of numerical error together. Suppose an exact solution $d_k$ has the form $d_k = -H_k^{-1}g_k$, then we define $\hat{d}_k$ is a $\delta$-approximate direction with respect to $d_k$ is a vector $\hat{d}_k$ satisfying

$$\|H_k\hat{d}_k + g_k\| \leq \delta. \tag{2.15}$$

In Algorithm 2.5 we propose the implicit Euler method with the presence of approximate direction of $v(\cdot, \cdot)$ at each iteration.

**Algorithm 2.5:** Approximate Euler method for solution path

**input :** Initial point $x_0 \in \mathbb{R}^p$, total number of iterations $K \geq 1$

Initialize regularization parameter $\lambda_0 \leftarrow \lambda_{\max}$;

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

  Update $\hat{d}_k \leftarrow$ an $\delta$-approximate direction of $v(x_k, \lambda_k)$;

  Update $\lambda_{k+1} \leftarrow (1-h)\lambda_k$;

  Update $x_{k+1} \leftarrow x_k + h \cdot \hat{d}_{k+1}$;

**output:** $\hat{x}(\cdot) \leftarrow \mathcal{I}_{\text{linear}}\left(\{(x_k, \lambda_k)\}_{k=0}^K\right)$ according to linear interpolation

Compared with the origin update scheme in Algorithm 2.2, the only difference in the update scheme of Algorithm 2.5 is that an approximate direction $\hat{d}_k$ is applied at each iteration instead of the exact direction $v(\cdot, \cdot)$. We want to mention that there is no constraint on how the approximate direction $\hat{d}_k$ is generated, and in Section 2.5.2, we provide several efficient methods to compute the approximate direction and the corresponding complexity analysis. The following lemma characterizes the local error accumulation of the update scheme in Algorithm 2.2.

**Lemma 2.5.1.** *Suppose Assumption 2.3.1 holds. Let $\hat{d}_k$ denote an approximate solution to $v(x_k, \lambda_k)$. Let $r_k = \|\nabla F_{\lambda_k}(x_k)\|$, $r_{k+1} = \|\nabla F_{\lambda_{k+1}}(x_{k+1})\|$, and*

$$\delta_k = \left(\nabla^2 f(x_k) + \lambda_{k+1}\nabla^2\Omega(x_k)\right)\hat{d}_k + \nabla f(x_k).$$

*Then it holds that*

$$r_{k+1} \leq \frac{\lambda_{k+1}}{\lambda_k} \cdot r_k + \frac{h^2 L(1+\lambda_{k+1})}{2} \cdot \|\hat{d}_k\|^2 + h\,\|\delta_k\|\,. \tag{2.16}$$

*Furthermore, if we set $\lambda_{s+1} = (1-h)\lambda_s$ and $\|\delta_s\| \leq \delta$ for some scalar $\delta > 0$ and all $s = 0, \ldots, k$, it holds that*

$$r_k \leq \frac{\lambda_k}{\lambda_0} \cdot r_0 + 2h\tau L(f(x_0) - f^*) + \delta. \tag{2.17}$$

*Proof.* The proof is similar with the one in Lemma 2.3.2. First we have

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

Then apply same technique as in Lemma 2.3.2 we arrive at

$$r_{k+1}\leq\frac{\lambda_{k+1}}{\lambda_k}\cdot r_k+h^2\cdot\frac{L(1+\lambda_{k+1})}{2}\|\hat{d}_k\|^2+h\left\|\delta_k\right\|.$$

Also, applying (2.16) to Proposition 2.3.2 implies the result in (2.17). □

The following corollary provides the complexity analysis of Algorithm 2.5.

**Corollary 2.5.1.** *Suppose that Assumption 2.3.1 holds, and suppose that the initial point $x_0$ satisfies $\|\nabla F_{\lambda_{\max}}(x_0)\|\leq\frac{\epsilon}{4}$. Let $T:=\log(\lambda_{\max}/\lambda_{\min})$, let $\tilde{\mu}:=\mu+\lambda_{\min}\sigma$, let $\epsilon\in(0,\tilde{\mu}]$ be the desired accuracy, and let*

$$K_{\text{E, approx}}:=\left\lceil\max\left\{2T,\frac{\sqrt{LG}\tau T}{\sqrt{3}},\frac{8(f(x_0)-f^*)\tau LT}{\epsilon},\frac{4\sqrt{L}(\tau(G+\epsilon)+1)T}{\sqrt{\epsilon}}\right\}\right\rceil.$$

*If the total number of iterations $K$ satisfies $K\geq K_{\text{E, approx}}$ and approximate directions $\hat{d}_k$ are all $\frac{\epsilon}{4}$-approximate, then Algorithm 2.5 returns an $\epsilon$-accurate solution path.*

*Proof.* Since the step-size $h=1-(\frac{\lambda_{\min}}{\lambda_{\max}})^{\frac{1}{K}}\leq\frac{T}{K}$, combining (2.17), we have $r_k\leq\frac{3\epsilon}{4}$, for all $k\in\{0,\ldots,K\}$. For linear interpolation error, we have

$$\begin{aligned}&\frac{L}{8}\cdot\max_{k\in[K-1]}\left\{(1+\lambda_k)\|x_{k+1}-x_k\|^2+2h|\xi(\lambda_k)|\|x_{k+1}-x_k\|\right\}\\\leq&\frac{h^2L}{2}\cdot\max_{k\in[K-1]}\left\{(1+\lambda_k)\left(\|d_k\|^2+\|d_k-\hat{d}_k\|^2\right)+|\xi(\lambda_k)|(\|d_k\|+\|d_k-\hat{d}_k\|)\right\}\\\leq&\frac{h^2L}{2}\cdot(\tau(G+\epsilon)+1)^2\leq\frac{\epsilon}{4}.\end{aligned}$$

Applying the above inequality to Theorem 2.3.2 completes the proof. □

Now we consider the trapezoid method with the presence of approximate direction and propose it in Algorithm 2.6.

**Algorithm 2.6:** Approximate trapezoid method for solution path

**input :** Initial point $x_0 \in \mathbb{R}^p$, total number of iterations $K \geq 1$

Initialize parameter $\lambda_0 \leftarrow \lambda_{\max}$, set step-size $h \leftarrow 1 - \sqrt{2(\frac{\lambda_{\min}}{\lambda_{\max}})^{\frac{1}{K}} - 1}$;

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

  Update $\hat{d}_{k,1} \leftarrow$ an $\delta$-approximate direction of $v(x_k, \lambda_k)$;

  Update $\hat{d}_{k,2} \leftarrow$ an $\delta$-approximate direction of $v(x_k + h\hat{d}_{k,1}, (1-h+h^2)\lambda_k)$;

  Update $\lambda_{k+1} \leftarrow (1 - h + \frac{h^2}{2})\lambda_k$;

  Update $x_{k+1} \leftarrow x_k + h \cdot \frac{\hat{d}_{k,1} + \hat{d}_{k,2}}{2}$;

**output:** $\hat{x}(\cdot) \leftarrow \mathcal{I}_{\text{linear}}\left(\{(x_k, \lambda_k)\}_{k=0}^{K}\right)$ according to linear interpolation

The update scheme in Algorithm 2.6 is similar as the one in Algorithm 2.3, but we apply approximate directions at each iteration in lieu of exact computation of $v(\cdot, \cdot)$. The following lemma characterize the local error accumulation of the update scheme in Algorithm 2.2.

**Lemma 2.5.2.** *Suppose $r_k = \|\nabla F_{\lambda_k}(x_k)\|$ satisfying that $r_k \leq \tilde{\mu}$. Let $\delta_{k,1}$ and $\delta_{k,2}$ denote the residual of approximate direction $\hat{d}_{k,1}$ and $\hat{d}_{k,2}$ and they satisfy the condition $\|\delta_{k,1}\|, \|\delta_{k,2}\| \leq \tilde{\mu}$. Then, it holds that*

$$r_{k+1} \leq \frac{\lambda_{k+1}}{\lambda_k} \cdot r_k + h^3 \cdot 3L(2+G)^3 + h^4 \cdot 2L^3\tau^2(2+G)^4 + \frac{h}{2}\|\delta_{k,1} - \delta_{k,2}\| + \frac{h^2}{2}\|\delta_{k,1}\|. \quad (2.18)$$

*Proof.* We will follow the idea in Lemmas 2.4.1, A.1.2 and A.1.3. Recall the result in Lemma A.1.2, 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 A.1.3 becomes

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

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 Lemma 2.4.1 to get (2.18). Then the right hand side of (A.4) becomes $-h\delta_1$. Also, the right hand side of (A.5) becomes $\frac{h}{2}(\delta_1 - \delta_2)$ and the right hand side of (A.6) becomes $\frac{h^2}{2}\lambda\nabla^2\Omega(x)d_1 + \frac{h^2}{2}\delta_1$. □

**Corollary 2.5.2.** *Suppose Assumption 2.4.1 holds, let $\tilde{\mu} := \mu + \lambda_{\min}\sigma$, let $\epsilon \in (0, \tilde{\mu}]$ be the desired accuracy, suppose that the initial point $x_0$ satisfies $\|\nabla F_{\lambda_{\max}}(x_0)\| \leq \frac{\epsilon}{4}$, let $T := 1.1\log(\lambda_{\max}/\lambda_{\min})$, and let*

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

*If the total number of iterations $K$ satisfies $K \geq K_{\text{tr}}$ and all approximate direction $\delta_{k,1}, \delta_{k,2}$ are $\frac{\epsilon}{2}$-approximate, then Algorithm 2.6 returns a $\epsilon$-accurate solution path.*

Corollaries 2.5.1 and 2.5.2 provide the complexity analysis of Euler method and trapezoid method with the presence of approximate directions. We can see that when the residual of approximate directions have a uniform upper bound over all iterations, Algorithms 2.5 and 2.6 have the complexity of the same order as Algorithms 2.2 and 2.3, which require exact directions. Moreover, Corollaries 2.5.1 and 2.5.2 only require $\epsilon$-approximate directions but no assumptions about how the approximate directions are generated. It provides flexibility in the choice of an approximate oracle to compute approximate directions.

### 2.5.2 Second-Order Conjugate Gradient Variants

Following the complexity analysis, we apply the conjugate gradient method to solve the subproblem, *i.e.*, to compute a $\delta$-approximate direction of $v(\cdot,\cdot)$. To measure the efficiency of second-order conjugate gradient type algorithms, we consider the computational complexity, *i.e.*, total calls to both gradient and Hessian-vector product oracle.

Now we apply the conjugate gradient method as an approximate oracle to compute the approximate direction $\hat{d}_k$ at each iteration. We use the approximate Euler method proposed in Algorithm 2.5 as an example. At iteration $k$, Algorithm 2.5 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 $g_k := \nabla f(x_k)$. We apply the conjugate gradient to approximately solve the equation $H_k\hat{d}_k + g_k = 0$ and set the initial guess to be the approximate direction $\hat{d}_{k-1}$ at the last iteration. Herein we provide the complexity analysis of Euler-CG method and trapezoid-CG method.

**Theorem 2.5.1.** *Suppose Assumption 2.3.1 holds, and suppose that the initial point $x_0$ satisfies $\|\nabla F_{\lambda_{\max}}(x_0)\| \leq \epsilon$. Let $T := \log(\lambda_{\max}/\lambda_{\min})$, let $\tilde{\mu} := \mu + \lambda_{\min}\sigma$, let $\epsilon \in (0, \tilde{\mu}]$ be the desired accuracy, and let*

$$K_{\text{E-cg}} \sim \tilde{\mathcal{O}}\left(\frac{L^{3/2}T(f(x_0) - f^*)}{\epsilon\tilde{\mu}^{3/2}}\right).$$

*If the total number of iterations $K$ satisfies $K \geq K_{\text{E-cg}}$, then Algorithm 2.5 via the conjugate gradient approximate oracle returns a $5\epsilon$-accurate solution path.*

Also, for the computational guarantee of approximate trapezoid method in Algorithm 2.6 via conjugate gradient approximate oracle we have similar argument.

**Theorem 2.5.2.** *Suppose Assumption 2.4.1 holds, and suppose that the initial point $x_0$ satisfies $\|\nabla F_{\lambda_{\max}}(x_0)\| \leq \epsilon$. Let $T := 1.1\log(\lambda_{\max}/\lambda_{\min})$, let $\tilde{\mu} := \mu + \lambda_{\min}\sigma$, let $\epsilon \in (0, \tilde{\mu}]$ be the desired accuracy, and let*

$$K_{\text{tr-cg}} \sim \tilde{\mathcal{O}}\left(\frac{L(1+G)^{3/2}T}{\epsilon^{1/2}\tilde{\mu}^{1/2}}\right).$$

*If the total number of iterations $K$ satisfies $K \geq K_{\text{tr-cg}}$, then Algorithm 2.6 via the conjugate gradient approximate oracle returns a $2\epsilon$-accurate solution path.*

*Proof of Theorems 2.5.1 and 2.5.2.* Here we only need to consider the inner complexity, *i.e.*, the number of iterations required to compute the approximate direction. For the Euler-CG algorithm, let $\{y_{k,s}\}$ denote the sequence generated by the conjugate gradient method with $y_{k,0} = \hat{d}_{k-1}$, let $H_k = \nabla^2 f(x_k) + \lambda_{k+1}\nabla^2\Omega(x_k)$, and let $g_k = \nabla f(x_k)$. Existing results of the conjugate gradient method guarantees that $\|H_k y_{k,s} + g_k\|_2 \leq 2\sqrt{\kappa_k}(1 - \frac{2}{\sqrt{\kappa_k}+1})^s \|H_k y_{k,0} + g_k\|_2$, where $\kappa_k$ is the condition number of $H_k$ and $\kappa_k \leq \frac{(1+\lambda_k)L}{\mu+\lambda_k\sigma} \leq \tau L$. Since the initial guess $y_0 = \hat{d}_{k-1}$ which is the approximate direction at last iteration, we have $\|H_{k-1}\hat{d}_{k-1} + g_{k-1}\| \leq \frac{\epsilon}{4}$. Then the initial guess guarantees that

$$\begin{aligned}\|H_k\hat{d}_{k-1} + g_k\|_2 &\leq \|H_{k-1}\hat{d}_{k-1} + g_{k-1}\|_2 + \|H_k\hat{d}_{k-1} + g_k - H_{k-1}\hat{d}_{k-1} - g_{k-1}\| \\ &\leq \frac{\epsilon}{4} + hL(1+\lambda_k)\left(\|\hat{d}_{k-1}\|^2 + \|\hat{d}_{k-1}\|\right) \leq \frac{\epsilon}{4} + 2hL(2+G)^2.\end{aligned}$$

Applying $h \sim \mathcal{O}(\frac{\epsilon}{(f(x_0)-f^*)\tau L})$, the inner complexity $N_k$ has an upper bound

$$N_k \leq \frac{\sqrt{\kappa_k}+1}{2}\log(\frac{2\sqrt{\kappa_k}\|H_k\hat{d}_{k-1} + g_k\|}{\epsilon/4}) \sim \tilde{\mathcal{O}}(\sqrt{\tau L}).$$

Therefore, the total computation complexity of Algorithm 2.5 with the conjugate gradient approximate oracle to compute an $\epsilon$-accurate solution path is $\tilde{\mathcal{O}}(\frac{L^{3/2}\tau^{3/2}(f(x_0)-f^*)T}{\epsilon})$. For the trapezoid-CG algorithm, let $x'_k = x_k + h\hat{d}_{k,1}$, $\lambda'_k = (1-h+h^2)\lambda_k$, $H_{k,1} := \nabla^2 f(x_k) + \lambda_k\nabla\Omega(x_k)$, $g_{k,1} := \nabla f(x_k)$, $H_{k,2} := \nabla^2 f(x'_k) + \lambda'_k\nabla\Omega(x'_k)$, and $g_{k,2} := \nabla f(x'_k)$. At iteration $k$, Algorithm 2.6 requires to compute the $\frac{\epsilon}{4}$ directions $\hat{d}_{k,1}$ and $\hat{d}_{k,2}$. In the conjugate gradient sub-routine, we use $\hat{d}_{k-1,1}$ as the warm-start for solving $\hat{d}_{k,1}$ and $\hat{d}_{k,1}$ for solving $\hat{d}_{k,2}$. Similarly, we have $\|H_{k,1}\hat{d}_{k-1,1} + g_{k,1}\| \leq \frac{\epsilon}{4} + 2hL(2+G)^2$ and $\|H_{k,2}\hat{d}_{k,1} + g_{k,2}\| \leq \frac{\epsilon}{4} + 2hL(2+G)^2$. Applying $h \sim \mathcal{O}(\frac{\epsilon^{1/2}}{L^{1/2}(2+G)^{3/2}})$, the inner complexity $N_{k,1}$ and $N_{k,2}$ have an upper bound $N_{k,1}, N_{k,2} \sim \tilde{\mathcal{O}}(\sqrt{\tau L})$. Therefore, the total computation complexity of Algorithm 2.6 with the conjugate gradient approximate oracle to compute an $\epsilon$-accurate solution path is $\tilde{\mathcal{O}}(\frac{L\tau^{1/2}(2+G)^{3/2}T}{\epsilon^{1/2}})$. □

Recall the complexity results in Remark 2.3.1, we notice that when we use the Nesterov's accelerated gradient method as the sub-problem solver, the total complexity will be $\frac{(\tau L)^{3/2}GT\log 2}{\epsilon}$. We can see the total complexity of Algorithm 2.6 has order $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ and is better than the best known in grid search method. Again, the complexity of the trapezoid method is better than the Euler method, since it exploit the higher-order smoothness of the function $f(\cdot)$ and $\Omega(\cdot)$.

## 2.6 Computational Experiments

In this section we present computational results of numerical experiments wherein we implement different version of discretization schemes and interpolation methods to compute the

approximate solution path. As a comparison with out method, we introduce two approach based on grid search methods proposed in [35, 63]. For sub-problem solver in the grid search methods, we use warm-started exact Newton method and Nesterov's accelerated gradient method to compare with our exact methods and SOCG variants. We focus on the following 8 versions of update schemes, where "CG" stands for conjugate gradient.

- Euler, Euler-CG: Algorithm 2.2, and Algorithm 2.5 using CG as the sub-problem approximate oracle.
- Trapezoid, Trapezoid-CG: Algorithm 2.3, and Algorithm 2.6 using CG as the sub-problem approximate oracle.
- Runge-Kutta, Runge Kutta-CG: Algorithm 2.4, and Algorithm 2.4 with approximate directions and using CG as the sub-problem approximate oracle.
- Grid Search-Newton/AGD: Algorithm use grid search method and the exact Newton/Nesterov's accelerated gradient method as the sub-problem solver.

### 2.6.1 Logistic Regression

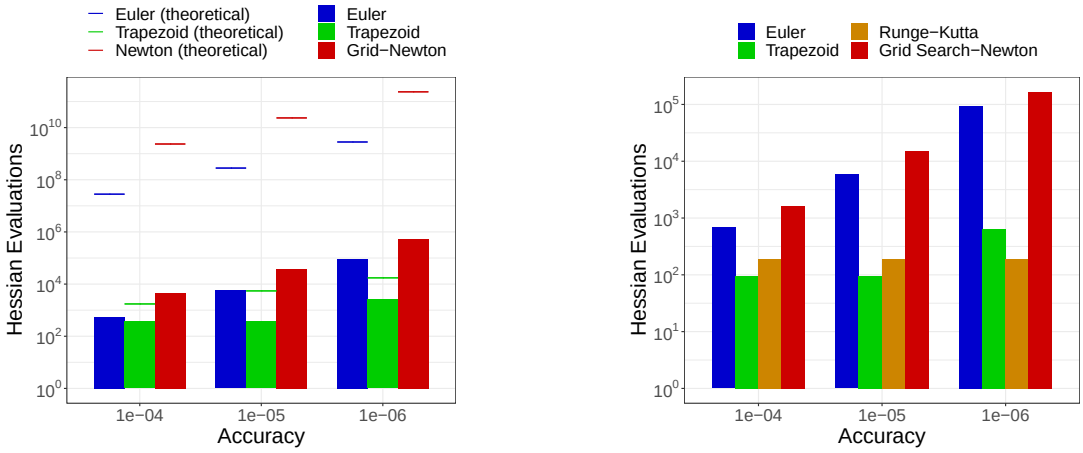
Herein we examine the empirical behavior of each of the previously presented methods on logistic regression problems using the breast cancer dataset from [26] (32 features and 569 observations) and the *leukemia* dataset from [37] (7129 features and 72 observations). In particular, let $\{(a_i, b_i)\}_{i=1}^n$ 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 logistic regression variants: *(i)* regularized logistic regression with

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-b_i a_i^T x}), \quad \Omega(x) = \frac{1}{2}\|x\|^2,$$

where $\lambda_{\min} = 10^{-4}$, $\lambda_{\max} = 10^4$, and *(ii)* re-weighted logistic regression with

$$f(x) = \frac{1}{|S_+|}\sum_{i\in S_+} \log(1 + e^{-b_i a_i^T x}), \quad \Omega(x) = \frac{1}{|S_-|}\sum_{i\in S_-} \log(1 + e^{-b_i a_i^T x}),$$

where $\lambda_{\min} = 10^{-1}$, $\lambda_{\max} = 10$. The initialization $x_0$ which we apply in each method is the same and is a very nearly-optimal solution to the problem such that $\|\nabla F_{\lambda_{\max}}(x_0)\| \approx 10^{-15}$. Note that it is hard to compute the exact path accuracy of an approximate solution path $\hat{x}(\cdot) : [\lambda_{\min}, \lambda_{\max}] \to \mathbb{R}^p$, namely $A(\hat{x}) := \max_{\lambda\in\Lambda} \|\nabla F_\lambda(\hat{x}(\lambda))\|$, where $\Lambda = [\lambda_{\min}, \lambda_{\max}]$. We examine the approximate path accuracy in lieu of exact computation of path accuracy, namely we consider $\hat{A}(\hat{x}) := \max_{\lambda\in\hat{\Lambda}} \|\nabla F_\lambda(\hat{x}(\lambda))\|$ where $\hat{\Lambda} = \{\lambda_0, \frac{\lambda_0+\lambda_1}{2}, \lambda_1, \frac{\lambda_1+\lambda_2}{2}, \lambda_2, \ldots, \lambda_K\}$, since the theoretical largest interpolation error occurs at the midpoint of two breakpoints. In Figure 2.1, we vary the desired accuracy parameter $\epsilon$ and plot the number of Hessian oracle computations required by the Algorithms 2.2 to 2.4, and the grid search method. The



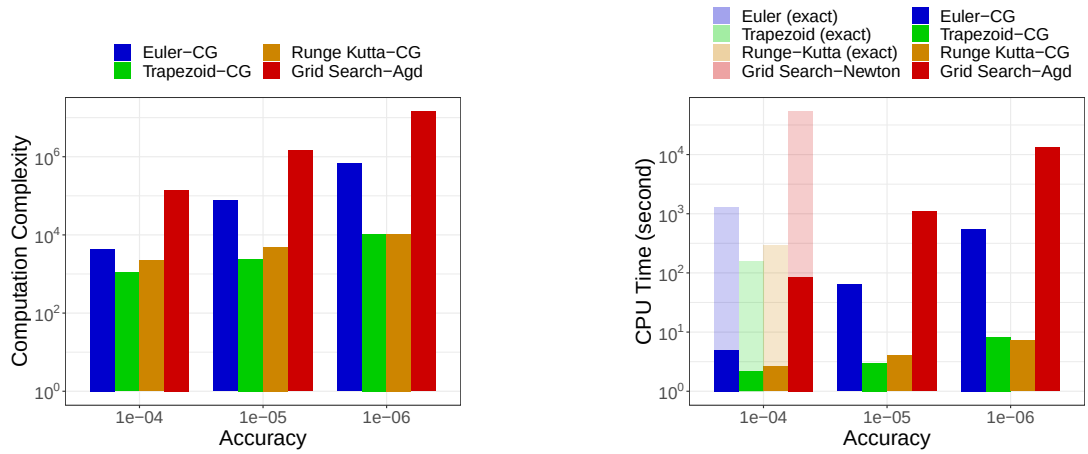

(a) Regularized logistic regression.

(b) Re-weighted logistic regression.

Figure 2.1: Exact methods on the breast cancer data with $n = 569$ observations and $p = 30$ features.

theoretical number is computed from the complexity analysis, and the practical number of each method is set according to a "doubling trick", whereby for each value of $K$ we calculate 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 $\epsilon$. The numerical results match the asymptotic order and the intuition, as well as the superior performance of the trapezoid and Runge-Kutta methods due to the higher-order smoothness of the loss and regularization function. In part *(a)* of Figure 2.1, we notice that the theoretical complexity is higher than the practical one since it is more conservative. Therefore, we will stick to the "doubling method" in the following experiments and compare the practical performance of each method.

Figure 2.2 summarizes the performance of the three aforementioned SOCG methods and the grid search method with Nesterov's accelerated gradient method. For SOCG methods, we record the total number of both gradient evaluation and Hessian-vector product, which is the computation complexity. We implement these methods on the *leukemia* dataset and the regularized logistic regression problem. From Figure 2.2, we can see that the numerical results match our theoretical asymptotic bounds. In part *(b)*, we provide the CPU time to compute the approximate solution of both SOCG methods and exact second-order methods (in the more transparent bars). We want to comment that for exact methods we only run the case when the desired accuracy equals to $10^{-4}$, since the grid search method already takes about 15 hours to finish. The dominance of SOCG methods in these cases illustrates the benefit and capability of SOCG methods to deal with large-scale problems.




(a) Computation complexity.

(b) CPU time.

Figure 2.2: Second-order conjugate gradient methods on regularized logistic regression on leukemia data with $n = 72$ observations and $p = 7129$ features.

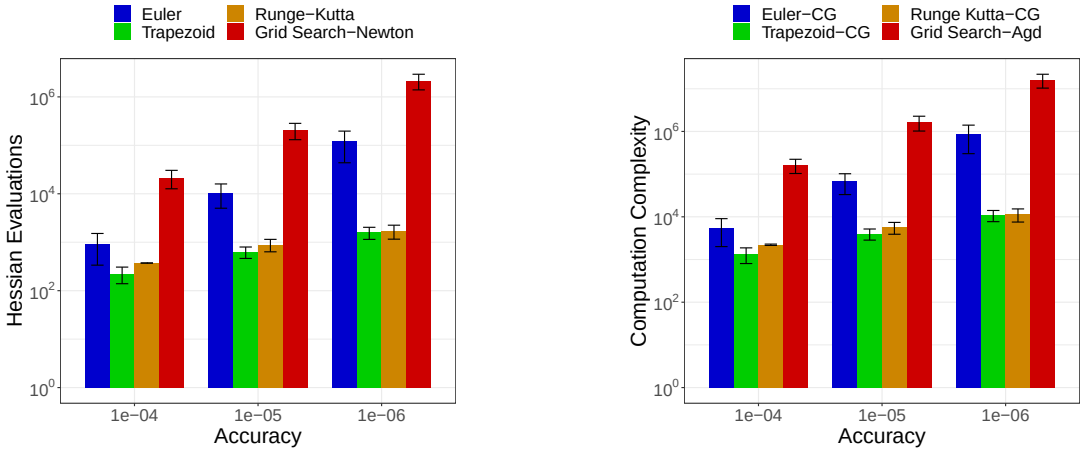
### 2.6.2 Moment Matching Problem

Herein we consider the moment matching problem with entropy regularization. Suppose a discrete random variable $Z$ has sample space $\{w_1, \ldots, w_{p+1}\}$ and probability distribution $\{x_1, \ldots, x_{p+1}\}$. We want to match the empirical first $n$-th moments of $Z$ with the entropy regularization. To formalize the problem, we consider the following constrained optimization problem:

$$\begin{aligned} P(\lambda): \quad \min_{x \in \mathbb{R}^{p+1}} \quad & \frac{1}{2}\|Ax - b\|^2 + \lambda \cdot \sum_{j=1}^{n} x_{(j)} \log(x_{(j)}) \\ \text{s.t.} \quad & \mathbf{1}_{p+1}^T x = 1, \quad x \geq 0, \end{aligned}$$

where $x_{(j)}$ is the $j$-th component of $x$, $A \in \mathbb{R}^{n \times (p+1)}$ with $A_{i,j} = w_j^i$, and $\lambda \in \Lambda = [10^{-2}, 10^2]$. The parametric optimization problem $P(\lambda)$ is a constrained optimization problem, which does not satisfy our setting. Therefore, we introduce a new variable $y$ to substitute $x$. Let $y \in \mathbb{R}^p$ with $y_{(i)} = x_{(i)}$, for $i = 1, \ldots, p$, and $S' = \{y \in \mathbb{R}^p : y \geq 0, \mathbf{1}_p^T y \leq 1\}$, it holds that $x \in S$ is equivalent to $y \in S'$ and $x \in \text{int}(S)$ is equivalent to $y \in \text{int}(S')$. We know that the moment matching problem $P(\lambda)$ is equivalent to the following parametric optimization problem:

$$\begin{aligned} P'(\lambda): \quad \min_{y \in \mathbb{R}^p} \quad & \frac{1}{2}\|A'y - b'\|^2 + \lambda \cdot \left( \sum_{j=1}^{n} y_{(j)} \log(y_{(j)}) + (1 - \mathbf{1}_p^T y) \log(1 - \mathbf{1}_p^T y) \right) \\ \text{s.t.} \quad & \mathbf{1}_p^T y \leq 1, \quad y \geq 0, \end{aligned}$$




(a) Exact methods.

(b) SOCG methods.

Figure 2.3: Exact and SOCG methods on moment matching problem with $n = 10$ and $p = 20$, the complexity comparison with different desired accuracy.

where $A' = A_{1:p} - A_{p+1}\mathbf{1}_p^T$ and $b' = b - A_{p+1}$. We examine the empirical behavior of each of the previously presented methods on $(P')$ without constraints, that is, $f(y) = \frac{1}{2}\|A'y - b'\|^2$ and $\Omega(y) = \sum_{j=1}^{n} y_{(j)} \log(y_{(j)}) + (1 - \mathbf{1}_p^T y) \log(1 - \mathbf{1}_p^T y)$. Lemma A.2.1 shows that the exact path $x(\lambda)$ for $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ is a subset of the relative interior of $S$. Also, when the step-size $h$ is small enough, all grid points $\{x_k\}$ will be in the relative interior of $S$, and therefore, the approximate path $\hat{x}(\lambda)$ for $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ is a subset of the relative interior of $S$.

**Synthetic Data Generation Process** We generate the data $(x, w)$ according to the following generative model. The true distribution vector $x^{\text{true}}$ is according to the model $x^{\text{true}}_{(i)} = \frac{\exp(z_{(i)})}{\sum_{j=1}^{p+1} \exp(z_{(j)})}$ for $i = 1, \ldots, p+1$, where $z_{(i)} \sim \text{unif}(0, 1)$. The sample vector $w$ are generated from a independent uniform distribution, *i.e.*, $w_{(i)} \sim \text{unif}(0, 1)$ for $i = 1, \ldots, p$, and without loss of generality we set $w_{(p+1)} = 0$.

First, we examine the aforementioned exact and SOCG methods with the different desired accuracy. Figure 2.3 summarizes our findings with $n = 10$ and $p = 20$, and the box plot with 95% confidence interval of each desired accuracy is across 10 independent trails. Again, Figure 2.3 demonstrates that the numerical results match the asymptotic order and the intuition, as well as the superior performance of the trapezoid and Runge-Kutta methods due to the higher-order smoothness of the loss and regularization function.

Moreover, we examine the aforementioned exact and SOCG methods with the different problem dimensions. In the following set of experiments, we set the number of observations $n = 20$, the desired accuracy $\epsilon = 10^{-5}$, and vary the problem dimension $p \in \{128, 256, 512, 1024, 2048\}$. The true distribution $x^{\text{true}}$ and the sample vector $w$ are synthet-

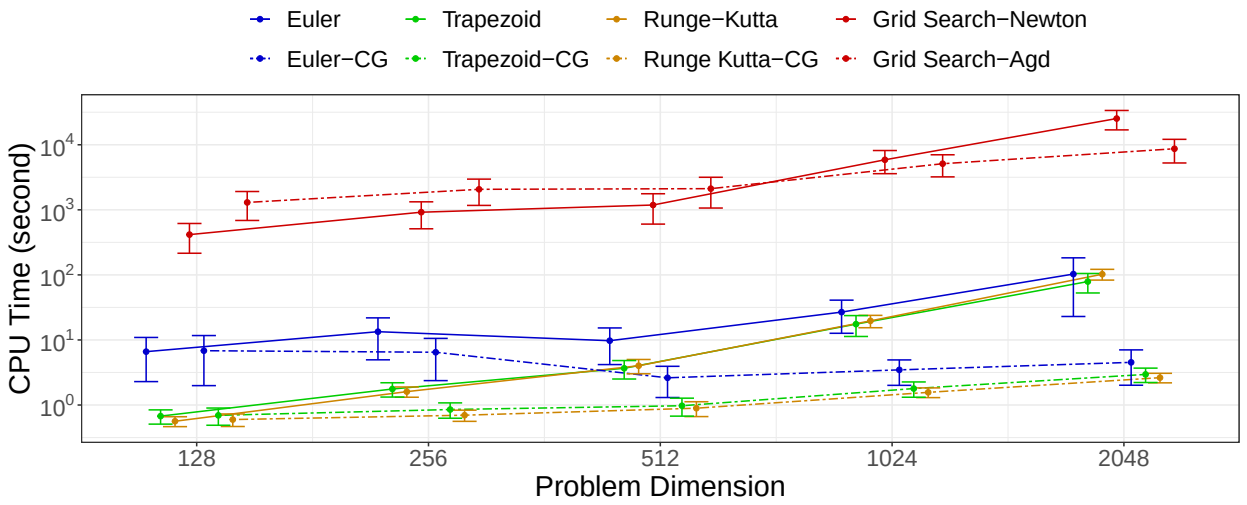



Figure 2.4: Exact and SOCG methods on moment matching problem with $n = 20$ and $\epsilon = 10^{-5}$, the CPU time comparison with different problem dimension.

ically generated according to the same process as before. Figure 2.4 displays our result for this experiment. Generally, we observe that the CPU time of computing an $\epsilon$-path increases as the problem dimension becomes larger. Comparing the CPU time of the exact methods and the SOCG methods, we notice that the SOCG methods are less sensitive to the sizes of the problems. Again, the results demonstrate the superiority of the SOCG methods to tackle large-scale problems.

# Chapter 3

# Risk Bounds and Calibration for a Smart Predict-then-Optimize Method

## 3.1 Introduction

The *predict-then-optimize* framework, where one predicts the unknown parameters of an optimization model and then plugs in the predictions before solving, is prevalent in applications of machine learning. Some typical examples include predicting future asset returns in portfolio allocation problems and predicting the travel time on each edge of a network in navigation problems. In most cases, there are many contextual features available, such as time of day, weather information, financial and business news headlines, and many others, that can be leveraged to predict the unknown parameters and reduce uncertainty in the decision making problem. Ultimately, the goal is to produce a high quality prediction model that leads to a good decisions when implemented, such as a position that leads to a large return or a route that induces a small realized travel time. There has been a fair amount of recent work examining this paradigm and other closely related problems in data-driven decision making, such as the works of [20, 25, 29, 44, 30, 40, 68, 48], the references therein, and others.

In this work, we focus on the particular and important case where the optimization problem of interest has a linear objective with a known convex feasible region and where the contextual features are related to the coefficients of the linear objective. This case includes the aforementioned shortest path and portfolio allocation problems. In this context, [29] developed the Smart Predict-then-Optimize (SPO) loss function, which directly measures the regret of a prediction against the best decision in hindsight (rather than just prediction error, such as squared error). After the introduction of the SPO loss, recent work has studied its statistical properties, including generalization bounds of the SPO loss function in [28] and generalization and regret convergence rates in [41]. Moreover, since the SPO loss is not continuous nor convex in general [29], which makes the training of a prediction model computationally intractable, [29] introduced a novel convex surrogate loss, referred

to as the SPO+ loss. [29] highlight and prove several advantages of the SPO+ surrogate loss: *(i)* it still accounts for the downstream optimization problem when evaluating the quality of a prediction model (unlike prediction losses such as the squared $\ell_2$ loss), *(ii)* it has a desirable Fisher consistency property with respect to the SPO loss under mild conditions, and *(iii)* it often performs better than commonly considered prediction losses in experimental results. Unfortunately, although a desirable property of any surrogate loss in this context, Fisher consistency is not directly applicable when one only has available a finite dataset, which is always the case in practice, because it relies on full knowledge of the underlying distribution. Motivated thusly, it is desirable to develop *risk bounds* that allow one to translate an approximate guarantee on the risk of a surrogate loss function to a corresponding guarantee on the SPO risk. That is, risk bounds (and the related notion of calibration functions) answer the question: to what tolerance $\delta$ should the surrogate excess risk be reduced to in order to ensure that the excess SPO risk is at most $\epsilon$? Note that, with enough data, it is possible in practice to ensure a (high probability) bound on the excess surrogate risk through generalization and optimization guarantees.

The main goal of this work is to provide risk bounds for the SPO+ surrogate loss function. Our results, to the best of our knowledge, are the first risk bounds of the SPO+ loss, besides the analysis of the 1-dimensional scenario in [67]. Our results consider two cases for the structure of the feasible region of the optimization problem: *(i)* the case of a bounded polyhedron, and *(ii)* the case of a level set of a smooth and strongly convex function. In the polyhedral case, we prove that the risk bound of the SPO+ surrogate is $\mathcal{O}(\epsilon^2)$, where $\epsilon$ is the desired accuracy for the excess SPO risk. Our results hold under mild distributional assumptions that extend those considered in [29]. In the strongly convex level set case, we improve the risk bound of the SPO+ surrogate to $\mathcal{O}(\epsilon)$ by utilizing novel properties of such sets that we develop, namely stronger optimality guarantees and continuity properties. As a consequence of our analysis, we can leverage generalization guarantees for the SPO+ loss to obtain the first sample complexity bounds, with respect to the SPO risk, for the SPO+ surrogate under the two cases we consider. In Section 3.5, we present computational results that validate our theoretical findings. In particular we present results on entropy constrained portfolio allocation problems which, to the best of our knowledge, is the first computational study of predict-then-optimize problems for a strongly convex feasible region. Our results on portfolio allocation problems demonstrate the effectiveness of the SPO+ surrogate. We also present results for cost-sensitive multi-class classification that illustrate the benefits of faster convergence of the SPO risk in the case of strongly convex sets as compared to polyhedral ones.

Starting with binary classification, risk bounds and calibration have been previously studied in other machine learning settings. Pioneer works studying the properties of convex surrogate loss functions for the 0-1 loss include [89, 17, 58] and [80]. Works including [88, 82] and [70] have studied the consistency and calibration properties of multi-class classification problems, which can be considered as a special case of the predict-then-optimize framework [28]. Most related to the results presented herein is the work of [67], who study the uniform calibration properties of the squared $\ell_2$ loss, and the related work of [41], who also develop

fast sample complexity results for the SPO loss when using a squared $\ell_2$ surrogate.

### 3.1.1 Organization

This chapter is organized as follows. In Section 3.2, after formally reviewing the Predict-then-optimize framework and the SPO and SPO+ loss, we discuss some existing results and methods for deriving risk bounds via calibration. Section 3.3 contains the generalization bounds of the SPO+ loss as well as the risk bounds and sample complexity of the polyhedron case. Section 3.4 contains the risk bounds and sample complexity of the level set case. Section 3.5 provides the numerical experiments on portfolio allocation instances.

### 3.1.2 Notation

Let $\odot$ represent element-wise multiplication between two vectors. For any positive integer $m$, let $[m]$ denote the set $\{1, \ldots, m\}$. Let $I_p$ denote the $p$ by $p$ identity matrix for any positive integer $p$. For $\bar{c} \in \mathbb{R}^d$ and a positive semi-definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, let $\mathcal{N}(\bar{c}, \Sigma)$ denote the normal distribution $\mathbb{P}(c) = \frac{e^{-\frac{1}{2}(c-\bar{c})^T \Sigma^{-1}(c-\bar{c})}}{\sqrt{(2\pi)^d \det(\Sigma)}}$. We will make use of a generic given norm $\|\cdot\|$ on $w \in \mathbb{R}^d$, as well as its dual norm $\|\cdot\|_*$ which is defined by $\|c\|_* = \max_{w:\|w\| \leq 1} c^T w$. For a positive definite matrix $A$, we define the $A$-norm by $\|w\|_A := \sqrt{w^T A w}$. Also, we denote the diameter of the set $S \subseteq \mathbb{R}^d$ by $D_S := \sup_{w,w' \in S} \|w - w'\|_2$.

## 3.2 Predict-then-Optimize Framework and Preliminaries

We now formally describe the predict-then-optimize framework, which is widely prevalent in stochastic decision making problems. We assume that the problem of interest has a linear objective, but the cost vector of the objective, $c \in \mathcal{C} \subseteq \mathbb{R}^d$, is not observed when the decision is made. Instead, we observe a feature vector $x \in \mathcal{X} \subseteq \mathbb{R}^p$, which provides contextual information associated with $c$. Let $\mathbb{P}$ denote the underlying joint distribution of the pair $(x, c)$. Let $w$ denote the decision variable and assume that we have full knowledge of the feasible region $S \subseteq \mathbb{R}^d$, which is assumed to be non-empty, compact, and convex. When a feature vector $x$ is provided, the goal of the decision maker is to solve the *contextual stochastic optimization problem:*

$$\min_{w \in S} \mathbb{E}_{c \sim \mathbb{P}(\cdot|x)}[c^T w] = \min_{w \in S} \mathbb{E}_{c \sim \mathbb{P}(\cdot|x)}[c]^T w. \tag{3.1}$$

As demonstrated by (3.1), for linear optimization problems the predict-then-optimize framework relies on a prediction of the conditional expectation of the cost vector, namely $\mathbb{E}[c|x] = \mathbb{E}_{c \sim \mathbb{P}(\cdot|x)}[c]$. Let $\hat{c}$ denote a prediction of the conditional expectation, then the next step in the predict-then-optimize setting is to solve the deterministic optimization problem with the

cost vector $\hat{c}$, namely

$$P(\hat{c}): \quad \min_{w \in S} \hat{c}^T w. \tag{3.2}$$

Depending on the structure of the feasible region $S$, the optimization problem $P(\cdot)$ can represent linear programming, conic programming, and even (mixed) integer programming, for example. In any case, we assume that we can solve $P(\cdot)$ to any desired accuracy via either a closed-form solution or a solver. Let $w^*(\cdot) : \mathbb{R}^d \to S$ denote a particular optimization oracle for problem (3.2), whereby $w^*(\hat{c})$ is an optimal solution of $P(\hat{c})$. (We assume that the oracle is deterministic and ties are broken in an arbitrary pre-specified manner.)

In order to obtain a model for predicting cost vectors, namely a cost vector predictor function $g : \mathcal{X} \to \mathbb{R}^d$, we may leverage machine learning methods to learn the underlying distribution $\mathbb{P}$ from observed data $\{(x_1, c_1), \ldots, (x_n, c_n)\}$, which are assumed to be independent samples from $\mathbb{P}$. Most importantly, following (3.1), we would like to learn the conditional expectation and thus $g(x)$ can be thought of as an estimate of $\mathbb{E}[c|x]$. We follow a standard recipe for learning a predictor function $g$ where we specify a loss function to measure the quality of predictions relative to the observed realized cost vectors. In particular, for a loss function $\ell$, the value $\ell(\hat{c}, c)$ represents the loss or error incurred when the cost vector prediction is $\hat{c}$ (i.e., $\hat{c} = g(x)$) and the realized cost vector is $c$. Let $R_\ell(g; \mathbb{P}) := \mathbb{E}_{(x,c)\sim\mathbb{P}}[\ell(g(x), c)]$ denote the the risk of given loss function $\ell$ and let $R^*_\ell(\mathbb{P}) = \inf_{g'} R_\ell(g'; \mathbb{P})$ denote the optimal $\ell$-risk over all measurable functions $g'$. Also, let $\hat{R}^n_\ell(g) := \frac{1}{n}\sum_{i=1}^n \ell(g(x_i), c_i)$ denote the empirical $\ell$-risk. Most commonly used loss functions are based on directly measuring the prediction error, including the (squared) $\ell_2$ and the $\ell_1$ losses. However, these losses do not take the downstream optimization task nor the structure of the feasible region $S$ into consideration. Motivated thusly, one may consider a loss function that directly measures the decision error with respect to the optimization problem (3.2). [29] formalize this notion in our context of linear optimization problems with their introduction of the SPO (Smart Predict-then-Optimize) loss function, which is defined by

$$\ell_{\text{SPO}}(\hat{c}, c) := c^T w^*(\hat{c}) - c^T w^*(c),$$

where $\hat{c} \in \mathbb{R}^d$ is the predicted cost vector and $c \in \mathcal{C}$ is the realized cost vector. Due to the possible non-convexity and possible discontinuities of the SPO loss, [29] also propose a convex surrogate loss function, the SPO+ loss, which is defined as

$$\ell_{\text{SPO+}}(\hat{c}, c) := \max_{w \in S}\{(c - 2\hat{c})^T w\} + 2\hat{c}^T w^*(c) - c^T w^*(c).$$

Importantly, the SPO+ loss still accounts for the downstream optimization problem (3.2) and the structure of the feasible region $S$, in contrast to losses that focus only on prediction error. As discussed by [29], the SPO+ loss can be efficiently optimized via linear/conic optimization reformulations and with (stochastic) gradient methods for large datasets. [29] provide theoretical and empirical justification for the use of the SPO+ loss function, including a derivation through duality theory, promising experimental results on shortest path and portfolio optimization instances, and the following theorem which provides the Fisher consistency of the SPO+ loss.

**Theorem 3.2.1** ([29], Theorem 1)**.** *Suppose that the feasible region $S$ has a non-empty interior. For fixed $x \in \mathcal{X}$, suppose that the conditional distribution $\mathbb{P}(\cdot|x)$ is continuous on all of $\mathbb{R}^d$, is centrally symmetric around its mean $\bar{c} := \mathbb{E}_{c\sim\mathbb{P}(\cdot|x)}[c]$, and that there is a unique optimal solution of $P(\bar{c})$. Then, for all $\Delta \in \mathbb{R}^p$ it holds that*

$$\mathbb{E}_{c\sim\mathbb{P}(\cdot|x)}\left[\ell_{\text{SPO+}}(\bar{c}+\Delta, c) - \ell_{\text{SPO+}}(\bar{c}, c)\right] = \mathbb{E}_{c\sim\mathbb{P}(\cdot|x)}\left[(c+2\Delta)^T(w^*(c) - w^*(c+2\Delta))\right] \geq 0.$$

*Moreover, if $\Delta \neq 0$, then $\mathbb{E}_{c\sim\mathbb{P}(\cdot|x)}\left[\ell_{\text{SPO+}}(\bar{c}+\Delta) - \ell_{\text{SPO+}}(\bar{c})\right] > 0$.*

Notice that Theorem 3.2.1 holds for *arbitrary* $x \in \mathcal{X}$, i.e., it employs a nonparametric analysis as is standard in consistency and calibration results, whereby there are no constraints on the predicted cost vector associated with $x$. Under the conditions of Theorem 3.2.1, given any $x \in \mathcal{X}$, we know that the conditional expectation $\bar{c} = \mathbb{E}_{c\sim\mathbb{P}(\cdot|x)}[c]$ is the unique minimizer of the SPO+ risk. Furthermore, since $\bar{c}$ is also a minimizer of the SPO risk, it holds that the SPO+ loss function is Fisher consistent with respect to the SPO loss function, i.e., minimizing the SPO+ risk also minimizes the SPO risk. However, in practice, due to the fact that we have available only a finite dataset and not complete knowledge of the distribution $\mathbb{P}$, we cannot directly minimize the true SPO+ risk. Instead, by employing the use of optimization and generalization guarantees, we are able to approximately minimize the SPO+ risk. A natural question is then: does a low excess SPO+ risk guarantee a low excess SPO risk? More formally, we are primarily interested in the following questions: *(i)* for any $\epsilon > 0$, does there exist $\delta(\epsilon) > 0$ such that $R_{\text{SPO+}}(g;\mathbb{P}) - R^*_{\text{SPO+}}(\mathbb{P}) < \delta(\epsilon)$ implies that $R_{\text{SPO}}(g;\mathbb{P}) - R^*_{\text{SPO}}(\mathbb{P}) < \epsilon$?, and *(ii)* what is the largest such value of $\delta(\epsilon)$ that guarantees the above?

### 3.2.1 Excess Risk Bounds via Calibration

The notions of calibration and calibration functions provide a useful set of tools to answer the previous questions. We now review basic concepts concerning calibration when using a generic surrogate loss function $\ell$, although we are primarily interested in the aforementioned SPO+ surrogate. An excess risk bound allows one to transfer the conditional excess $\ell$-risk, $\mathbb{E}\left[\ell(\hat{c}, c)|x\right] - \inf_{c'}\mathbb{E}\left[\ell(c', c)|x\right]$, to the conditional excess $\ell_{\text{SPO}}$-risk, $\mathbb{E}\left[\ell_{\text{SPO}}(\hat{c}, c)|x\right] - \inf_{c'}\mathbb{E}\left[\ell_{\text{SPO}}(c', c)|x\right]$. Calibration, which we now briefly review, is a central tool in developing excess risk bounds. We adopt the definition of calibration presented by [80] and [67], which is reviewed in Definition 3.2.1 below.

**Definition 3.2.1.** *For a given surrogate loss function $\ell$, we say $\ell$ is $\ell_{\text{SPO}}$-calibrated with respect to $\mathbb{P}$ if there exists a function $\delta_\ell(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $x \in \mathcal{X}$, $\hat{c} \in \mathcal{C}$, and $\epsilon > 0$, it holds that*

$$\mathbb{E}\left[\ell(\hat{c}, c)|x\right] - \inf_{c'}\mathbb{E}\left[\ell(c', c)|x\right] < \delta_\ell(\epsilon) \Rightarrow \mathbb{E}\left[\ell_{\text{SPO}}(\hat{c}, c)|x\right] - \inf_{c'}\mathbb{E}\left[\ell_{\text{SPO}}(c', c)|x\right] < \epsilon. \tag{3.3}$$

*Additionally, if (3.3) holds for all $\mathbb{P} \in \mathcal{P}$, where $\mathcal{P}$ is a class of distributions on $\mathcal{X} \times \mathcal{C}$, then we say that $\ell$ is uniformly calibrated with respect to the class of distributions $\mathcal{P}$.*

A direct approach to finding a feasible $\delta_\ell(\cdot)$ function and checking for uniform calibration is by computing the infimum of the excess surrogate loss subject to a constraint that the excess SPO loss is at least $\epsilon$. This idea leads to the definition of the *calibration function*, which we review in Definition 3.2.2 below.

**Definition 3.2.2.** *For a given surrogate loss function $\ell$ and true cost vector distribution $\mathbb{P}_c$, the conditional calibration function $\hat{\delta}_\ell(\cdot; \mathbb{P}_c)$ is defined, for $\epsilon > 0$, by*

$$\hat{\delta}_\ell(\epsilon; \mathbb{P}_c) := \inf_{\hat{c} \in \mathbb{R}^d} \left\{ \mathbb{E}\left[\ell(\hat{c}, c)\right] - \inf_{c'} \mathbb{E}\left[\ell(c', c)\right] \;\; : \;\; \mathbb{E}\left[\ell_{\text{SPO}}(\hat{c}, c)\right] - \inf_{c'} \mathbb{E}\left[\ell_{\text{SPO}}(c', c)\right] \geq \epsilon \right\}.$$

*Moreover, given a class of joint distributions $\mathcal{P}$, with a slight abuse of notation, the calibration function $\hat{\delta}_\ell(\cdot; \mathcal{P})$ is defined, for $\epsilon > 0$, by*

$$\hat{\delta}_\ell(\epsilon; \mathcal{P}) := \inf_{x \in \mathcal{X}, \mathbb{P} \in \mathcal{P}} \hat{\delta}_\ell(\epsilon; \mathbb{P}(\cdot|x)).$$

If the calibration function $\hat{\delta}_\ell(\cdot; \mathcal{P})$ satisfies $\hat{\delta}_\ell(\epsilon; \mathcal{P}) > 0$ for all $\epsilon > 0$, then the loss function $\ell$ is uniformly $\ell_{\text{SPO}}$-calibrated with respect to the class of distributions $\mathcal{P}$. To obtain an excess risk bound, we let $\delta_\ell^{**}$ denote the biconjugate, the largest convex lower semi-continuous envelope, of $\delta_\ell$. Jensen's inequality then readily yields $\delta_\ell^{**}(R_{\text{SPO}}(g, \mathbb{P}) - R^*_{\text{SPO}}(\mathbb{P})) \leq R_\ell(g, \mathbb{P}) - R_\ell(\mathbb{P})$, which implies that the excess surrogate risk $R_\ell(g, \mathbb{P}) - R_\ell(\mathbb{P})$ of a predictor $g$ can be translated into an upper bound of the excess SPO risk $R_{\text{SPO}}(g, \mathbb{P}) - R^*_{\text{SPO}}(\mathbb{P})$. For example, the uniform calibration of the least squares (squared $\ell_2$) loss, namely $\ell_{\text{LS}}(\hat{c}, c) = \|\hat{c} - c\|_2^2$, was examined by [67]. They proved that the calibration function is $\delta_{\ell_{\text{LS}}}(\epsilon) = \epsilon^2 / D_S^2$, which implies an upper bound of the excess SPO risk by $R_{\text{SPO}}(g, \mathbb{P}) - R^*_{\text{SPO}}(\mathbb{P}) \leq D_S (R_{\text{LS}}(g, \mathbb{P}) - R^*_{\text{LS}}(\mathbb{P}))^{1/2}$. In this paper, we derive the calibration function of the SPO+ loss and thus reveal the quantitative relationship between the excess SPO risk and the excess surrogate SPO+ risk under different circumstances.

### 3.2.2 Rademacher Complexity and Generalization Bounds

Herein we briefly review *Rademacher complexity*, a widely used concept in deriving generalization bounds, and how it applies in our analysis. For any loss function $\ell(\cdot, \cdot)$ and a hypothesis class $\mathcal{H}$ of cost vector predictor functions, the Rademacher complexity is defined as

$$\mathfrak{R}_\ell^n(\mathcal{H}) := \mathbb{E}_{\sigma, \{(x_i, c_i)\}_{i=1}^n} \left[ \sup_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(g(x_i), c_i) \right],$$

where $\sigma_i$ are independent Rademacher random variables and $(x_i, c_i)$ are independent samples from the joint distribution $\mathbb{P}$ for $i = 1, \ldots, n$. The following theorem provides a classical generalization bounds based on the Rademacher complexity.

**Theorem 3.2.2** ([18]). *Let $\mathcal{H}$ be a hypothesis class from $\mathcal{X}$ to $\mathbb{R}^d$ and let $b = \sup_{\hat{c}\in\mathcal{H}(\mathcal{X}),c\in\mathcal{C}} \ell(\hat{c},c)$. Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $g \in \mathcal{H}$ it holds that*

$$\left|R_\ell(g;\mathbb{P}) - \hat{R}^n_\ell(g)\right| \leq 2\mathfrak{R}^n_\ell(\mathcal{H}) + b\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Moreover, we define the *multivariate Rademacher complexity* [59, 20, 28] of $\mathcal{H}$ as

$$\mathfrak{R}^n(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma},x}\left[\sup_{g\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\sigma}_i^T g(x_i)\right],$$

where $\boldsymbol{\sigma}_i \in \{-1,+1\}^d$ are Rademacher random vectors for $i = 1,\ldots,n$. In many cases of hypothesis classes, such as linear functions with bounded Frobenius or element-wise $\ell_1$ norm, the multivariate Rademacher complexity can be bounded as $\mathfrak{R}^n(\mathcal{H}) \leq \frac{C'}{\sqrt{n}}$ where $C'$ is a constant that usually depends on the properties of the data, the hypothesis class, and mildly on the dimensions $d$ and $p$. Detailed examples of such bounds have been provided by [28, 20].

When the loss function $\ell(\cdot,\cdot)$ is additionally $L$-Lipschitz continuous with respect to the 2-norm in the first argument, namely $|\ell(\hat{c}_1,c) - \ell(\hat{c}_2,c)| \leq L\|\hat{c}_1 - \hat{c}_2\|_2$ for all $\hat{c}_1,\hat{c}_2,c \in \mathbb{R}^p$, then by the vector contraction inequality of [59] we have $\mathfrak{R}^n_\ell(\mathcal{H}) \leq \sqrt{2}L\mathfrak{R}^n(\mathcal{H})$. It is also easy to see that the the SPO+ loss function $\ell_{\text{SPO+}}(\cdot,c)$ is $2D_S$-Lipschitz continuous with respect to the 2-norm for any $c$ and therefore we can leverage the vector contraction inequality of [59] in this case. Combined with Theorem 3.2.2, this yields a generalization bound for the SPO+ loss.

## 3.3 Risk Bounds and Calibration for Polyhedral Sets

In this section, we consider the case when the feasible region $S$ is a bounded polyhedron and derive the calibration function of the SPO+ loss function. As is shown in Theorem 3.2.1, the SPO+ loss is Fisher consistent when the conditional distribution $\mathbb{P}(\cdot|x)$ is continuous on all of $\mathbb{R}^d$ and is centrally symmetric about its mean $\bar{c}$. More formally, the joint distribution $\mathbb{P}$ lies in the distribution class $\mathcal{P}_{\text{cont, symm}} := \{\mathbb{P} : \mathbb{P}(\cdot|x)$ is continuous on all of $\mathbb{R}^d$ and is centrally symmetric about its mean, for all $x \in \mathcal{X}\}$. In Example 3.3.2, we later show that this distribution class is not restrictive enough to obtain a meaningful calibration function. Instead, we consider a more specific distribution class consisting of distributions whose density functions can be lower bounded by a normal distribution. More formally, for given parameters $M \geq 1$ and $\alpha,\beta,D > 0$, define $\mathcal{P}_{M,\alpha,\beta,D} := \{\mathbb{P} \in \mathcal{P}_{\text{cont, symm}} :$ for all $x \in \mathcal{X}$ with $\bar{c} = \mathbb{E}[c|x]$, there exists $\sigma \in [0,\min\{D,M\}]$ satisfying $\|\bar{c}\|_2 \leq \beta\sigma$ and $\mathbb{P}(c|x) \geq \alpha\cdot\mathcal{N}(\bar{c},\sigma^2 I)$ for all $c \in \mathbb{R}^d$ satisfying $\|c\|_2^2 \leq 2D^2\}$. Also, let $\mathcal{P}_{M,\alpha,\beta} \leftarrow \mathcal{P}_{M,\alpha,\beta,\infty}$. Intuitively, the assumptions on the distribution class $\mathcal{P}_{M,\alpha,\beta,D}$ ensure that we avoid a situation where the density of the cost vector concentrates around some "badly behaved points." This intuition is further highlighted in Example 3.3.2. Theorem 3.3.1 is our main result in the polyhedral

case and demonstrates that the previously defined distribution class is a sufficient class to obtain a positive calibration function. Recall that $D_S$ denotes the diameter of $S$ and define a "width constant" associated with $S$ by $d_S := \min_{v\in\mathbb{R}^d:\|v\|_2=1}\left\{\max_{w\in S} v^T w - \min_{w\in S} v^T w\right\}$. Notice that $d_S > 0$ whenever $S$ has a non-empty interior.

**Theorem 3.3.1.** *Suppose the feasible region $S$ is a polyhedron and define $\Xi_S := (1 + \frac{2\sqrt{3}D_S}{d_S})^{1-d}$. Then the calibration function of the SPO+ loss satisfies*

$$\hat{\delta}_{\ell_{\text{SPO+}}}(\epsilon; \mathcal{P}_{M,\alpha,\beta,D}) \geq \frac{\alpha\Xi_S\gamma(\frac{d-1}{2}, D^2)}{4\sqrt{2\pi}e^{\frac{3(1+\beta^2)}{2}}\Gamma(\frac{d-1}{2})} \cdot \min\left\{\frac{\epsilon^2}{D_S M}, \epsilon\right\} \quad \textit{for all } \epsilon > 0. \tag{3.4}$$

*Additionally, when $D = \infty$, we have $\gamma(\frac{d-1}{2}, D^2) = \Gamma(\frac{d-1}{2})$ and therefore*

$$\hat{\delta}_{\ell_{\text{SPO+}}}(\epsilon; \mathcal{P}_{M,\alpha,\beta}) \geq \frac{\alpha\Xi_S}{4\sqrt{2\pi}e^{\frac{3(1+\beta^2)}{2}}} \cdot \min\left\{\frac{\epsilon^2}{D_S M}, \epsilon\right\} \quad \textit{for all } \epsilon > 0. \tag{3.5}$$

Theorem 3.3.1 yields an $\mathcal{O}(\epsilon^2)$ uniform calibration result for the distribution class $\mathcal{P}_{M,\alpha,\beta,D}$. The dependence on the constants is also natural as it matches the upper bound given by the cases with a $\ell_1$-like unit ball feasible region $S$ and standard multivariate normal distribution as the conditional probability $\mathbb{P}(\cdot|x)$. The following example shows the tightness of the lower bound in Theorem 3.3.1.

**Example 3.3.1.** *For any given $\epsilon > 0$, we consider the conditional distribution $\mathbb{P}(c|x) = \mathcal{N}(-\epsilon' \cdot e_d, \sigma^2 I_d)$ for some constants $\epsilon', \sigma > 0$ to be determined. For some $a, b > 0$, let the feasible region $S$ be $S = \text{conv}(\{w \in \mathbb{R}^d : \|w_{1:(d-1)}\|_2 = a, w_d = 0\} \cup \{\pm b \cdot e_d\})$. Although $S$ is not polyhedral, it can be considered as a limiting case of a polyhedron and the argument easily extends, with minor complications, to the case where the sphere is replaced by an $(d-1)$-gon for $d$ sufficiently large. Let $\hat{c} = \epsilon' \cdot e_d$, we have $\mathbb{E}\left[\ell_{\text{SPO}}(\hat{c}, c)|x\right] - \mathbb{E}\left[\ell_{\text{SPO}}(\bar{c}, c)|x\right] = 2b\epsilon'$. Also, for the excess conditional SPO+ risk we have*

$$\begin{aligned}
\mathbb{E}\left[\ell_{\text{SPO+}}(\hat{c}, c)|x\right] - \mathbb{E}\left[\ell_{\text{SPO+}}(\bar{c}, c)|x\right] &\to \int_{\mathbb{R}^{d-1}} \prod_{j=1}^{d-1} \frac{e^{-\frac{c_j^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \cdot \frac{e^{-\frac{a^2\sum_{j=1}^{d-1} c_j^2}{2b^2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \cdot \frac{\epsilon'^2}{2} \cdot \mathrm{d}c_1 \ldots \mathrm{d}c_{d-1} \\
&= \frac{\epsilon'^2}{2\sqrt{2\pi\sigma^2}} \prod_{j=1}^{d-1} \int_{\mathbb{R}} \frac{e^{-\frac{c_j^2}{2\sigma^2}} \cdot e^{-\frac{a^2 c_j^2}{2b^2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \mathrm{d}c_j \\
&= \frac{\epsilon'^2}{2\sqrt{2\pi\sigma^2}} \cdot \left(\frac{b^2}{a^2+b^2}\right)^{(d-1)/2},
\end{aligned}$$

*when $\epsilon' \to 0$. Therefore, let $\epsilon' = \frac{\epsilon}{2b}$, we have $\mathbb{E}\left[\ell_{\text{SPO}}(\hat{c}, c)|x\right] - \mathbb{E}\left[\ell_{\text{SPO}}(\bar{c}, c)|x\right] = \epsilon$ and*

$$\begin{aligned}\mathbb{E}\left[\ell_{\text{SPO+}}(\hat{c}, c)|x\right] - \mathbb{E}\left[\ell_{\text{SPO+}}(\bar{c}, c)|x\right] &= \frac{\epsilon^2}{8\sqrt{2\pi\sigma^2 b^2}} \cdot \left(\frac{b^2}{a^2 + b^2}\right)^{(d-1)/2} \\ &\leq \frac{1}{8\sqrt{2\pi}} \cdot \left(\frac{D_S}{d_S}\right)^{1-d} \cdot \frac{\epsilon^2}{\sigma},\end{aligned}$$

*for some $b$ large enough, and therefore the lower bound in Theorem 3.3.1 is tight up to a constant.*

Let us now provide some more intuition on the parameters involved in the definition of the distribution class $\mathcal{P}_{M,\alpha,\beta,D}$ and their roles in Theorem 3.3.1. In the definition of $\mathcal{P}_{M,\alpha,\beta,D}$, $\alpha$ is a lower bound on the ratio of the density of the distribution of the cost vector relative to a "reference" standard normal distribution. When $\alpha$ is larger, the distribution is behaved more like a normal distribution and it leads to a better lower bound on the calibration function (3.4) and (3.5) in Theorem 3.3.1. The parameter $M$ is an upper bound on the standard deviation of the aforementioned reference normal distribution, and the lower bound (3.4) and (3.5) naturally become worse as $M$ increases. The parameter $\beta$ measures how the conditional mean deviates from zero relative to the standard deviation of the reference normal distribution. If this distance is larger then the predictions are larger on average, and the bounds in (3.4) and (3.5) become worse. The width constant $d_S$ measures the near-degeneracy of the polyhedron ($d_S = 0$ is degenerate) and the bound becomes meaningless as $d_S \to 0$. When the feasible region $S$ is near-degenerate, i.e., the ratio $\frac{d_S}{D_S}$ is close to zero, we tend to have a weaker lower bound on the calibration function, which is also natural.

**Example 3.3.2.** *Let the feasible region be the $\ell_1$ ball $S = \{w \in \mathbb{R}^2 : \|w\|_1 \leq 1\}$ and consider the distribution class $\mathcal{P}_{\text{cont, symm}}$. For a fixed scalar $\epsilon > 0$, let $c_1 = (9\epsilon, 0)^T$ and $c_2 = (-7\epsilon, 0)^T$. Let the conditional distribution $\mathbb{P}_\sigma(c|x)$ be a mixture of Gaussians defined by $\mathbb{P}_\sigma(c|x) := \frac{1}{2}(\mathcal{N}(c_1, \sigma^2 I) + \mathcal{N}(c_2, \sigma^2 I))$ for any $\sigma > 0$, and we have $\mathbb{P}_\sigma(c|x) \in \mathcal{P}_{\text{cont, symm}}$. Let the predicted cost vector be $\hat{c} = (0, \epsilon)^T$, then the excess conditional SPO risk is $\mathbb{E}[\ell_{\text{SPO}}(\hat{c}, c) - \ell_{\text{SPO}}(\bar{c}, c)|x] = \epsilon$. Then it holds that the excess conditional SPO+ risk $\mathbb{E}\left[\ell_{\text{SPO+}}(\hat{c}, c) - \ell_{\text{SPO+}}(\bar{c}, c)|x\right] \to 0$ when $\sigma \to 0$, and hence we have $\hat{\delta}_\ell(\epsilon; \mathcal{P}_{\text{cont, symm}}) = 0$.*

The intuition of Example 3.3.2 is that the existence of a non-zero calibration function requires the conditional distribution of $c$ given $x$ to be "uniform" on the space $\mathbb{R}^d$, but not concentrate near certain points. Example 3.3.2 highlights a situation that considers one such "badly behaved" case where a limiting distribution of a mixture of two Gaussians leads to a zero calibration function.

**Detailed Derivation for Example 3.3.2** Let the feasible region be the $\ell_1$ ball $S = \{w \in \mathbb{R}^2 : \|w\|_1 \leq 1\}$ and consider the distribution class $\mathcal{P}_{\text{cont, symm}}$. Let $x \in \mathcal{X}$ be fixed, $\epsilon > 0$ be a fixed scalar, $c_1 = (9\epsilon, 0)^T$ and $c_2 = (-7\epsilon, 0)^T$. Let the conditional distribution be a mixture of normals defined by $\mathbb{P}_\sigma(c|x) := \frac{1}{2}(\mathcal{N}(c_1, \sigma^2 I) + \mathcal{N}(c_2, \sigma^2 I))$ for some $\sigma > 0$. The

condition mean of $c$ is then $\bar{c} = (\epsilon, 0)^T$ and the distribution $\mathbb{P}_\sigma(c|x)$ is centrally symmetric around $\bar{c}$; therefore $\mathbb{P}_\sigma \in \mathcal{P}_{\text{cont, symm}}$. Let $\hat{c} = (0, \epsilon)^T$ and $\Delta := \hat{c} - \bar{c}$, which yields that the excess conditional SPO risk is $\mathbb{E}[\ell_{\text{SPO}}(\hat{c}, c) - \ell_{\text{SPO}}(\bar{c}, c)] = \bar{c}^T(w^*(\hat{c}) - w^*(\bar{c})) = \epsilon$. Also, for all $c \in \mathcal{C}$, we may assume that $w^*(c) \in Z_S = \{\pm e_1, \pm e_2\}$ and hence $(c + 2\Delta)^T(w^*(c) - w^*(c + 2\Delta)) \leq 2\Delta^T(w^*(c) - w^*(c + 2\Delta)) \leq 4\epsilon$. Therefore, using $\mathbb{E}\left[\ell_{\text{SPO+}}(\bar{c} + \Delta, c) - \ell_{\text{SPO+}}(\bar{c}, c)\right] = \mathbb{E}\left[(c + 2\Delta)^T(w^*(c) - w^*(c + 2\Delta))\right]$, it holds that

$$\begin{aligned}\mathbb{E}\left[\ell_{\text{SPO+}}(\bar{c} + \Delta, c) - \ell_{\text{SPO+}}(\bar{c}, c)\right] &\leq 4\epsilon\mathbb{P}_\sigma(w^*(c) \neq w^*(c + 2\Delta)) \\ &\leq 4\epsilon(1 - \mathbb{P}_\sigma(\{\|c - c_1\|_2 \leq \epsilon\} \cup \{\|c - c_2\|_2 \leq \epsilon\})) \to 0,\end{aligned}$$

when $\sigma \to 0$, and hence we have $\hat{\delta}_\ell(\epsilon; \mathcal{P}_{\text{cont, symm}}) = 0$.

By combining Theorem 3.3.1 with a generalization bound for the SPO+ loss, we can develop a sample complexity bound with respect to the SPO loss. Corollary 3.3.1 below presents such a result for the SPO+ method with a polyhedral feasible region. The derivation of Corollary 3.3.1 relies on the notion of *multivariate Rademacher complexity* as well as the vector contraction inequality of [59] in the $\ell_2$-norm. In particular, for a hypothesis class $\mathcal{H}$ of cost vector predictor functions (functions from $\mathcal{X}$ to $\mathbb{R}^d$), the multivariate Rademacher complexity is defined as $\mathfrak{R}^n(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma},x}\left[\sup_{g\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\sigma}_i^T g(x_i)\right]$, where $\boldsymbol{\sigma}_i \in \{-1, +1\}^d$ are Rademacher random vectors for $i = 1, \ldots, n$. Please refer to Section 3.2.2 for a detailed discussion of multivariate Rademacher complexity and the derivation of Corollary 3.3.1.

**Corollary 3.3.1.** *Suppose that the feasible region $S$ is a bounded polyhedron, the optimal predictor $g^*(x) = \mathbb{E}[c|x]$ is in the hypothesis class $\mathcal{H}$, and there exists a constant $C'$ such that $\mathfrak{R}^n(\mathcal{H}) \leq \frac{C'}{\sqrt{n}}$. Let $\hat{g}^n_{\text{SPO+}}$ denote the predictor which minimizes the empirical SPO+ risk $\hat{R}^n_{\text{SPO+}}(\cdot)$ over $\mathcal{H}$. Then there exists a constant $C$ such that for any $\mathbb{P} \in \mathcal{P}_{M,\alpha,\beta}$ and $\delta \in (0, \frac{1}{2})$, with probability at least $1 - \delta$, it holds that*

$$R_{\text{SPO}}(\hat{g}^n_{\text{SPO+}}; \mathbb{P}) - R^*_{\text{SPO}}(\mathbb{P}) \leq \frac{C\sqrt{\log(1/\delta)}}{n^{1/4}}.$$

The proof is provided after Corollary 3.4.1. Notice that the rate in Corollary 3.3.1 is $\mathcal{O}(1/n^{1/4})$. However, the bound is with respect to the SPO loss which is generally non-convex and is the first such bound for the SPO+ surrogate. [41] present a similar result for the squared $\ell_2$ surrogate with a rate of $\mathcal{O}(1/\sqrt{n})$, and an interesting open question concerns whether the rate can also be improved for the SPO+ surrogate.

In the remaining part of this section, we provide the proof of Theorem 3.3.1 and some useful lemmas.

**Additional Definitions and Notation.** Recall that $S$ is polyhedral and let $Z_S$ denote the extreme points of $S$. We assume, for simplicity, that $w^*(c) \in Z_S$ for all $c \in \mathbb{R}^d$, but our results can be extended to allow for other possibilities in the case when there are multiple optimal solutions of $P(c)$. For any $i \in \{1, \ldots, d\}$, we use $e_i \in \mathbb{R}^d$ to represent the unit vector whose

$i$-th entry is 1 and others are all zero. Given a vector $c' \in \mathbb{R}^{d-1}$ and a scalar $\xi \in \mathbb{R}$, let $(c', \xi)$ denote the vector $(c'^T, \xi)^T \in \mathbb{R}^d$. For fixed $c'$ and when $\xi$ ranges from negative infinity to positive infinity, the corresponding optimal solution $w^*(c', \xi)$ will sequentially take different values in $Z_S$, and we let $\Omega(c') = (w_1(c'), \ldots, w_{k(c')}(c'))$ denote this sequence. Let $y_i(c')$ denote the last element of vector $w_i(c')$ for $i = 1, \ldots, k(c')$. Also, for $i = 1, \ldots, k(c') - 1$, we define phase transition location $\zeta_i(c') \in \mathbb{R}$ such that $(c', \zeta_i(c'))^T w_i(c') = (c', \zeta_i(c'))^T w_{i+1}(c')$, and additionally, we define $\zeta_0(c') = -\infty$ and $\zeta_{k(c')}(c') = \infty$. When there is no confusion, we will omit $c'$ and only use $k, w_i, y_i, \zeta_i$ for simplicity. Based on the above definition, for all $\xi \in (\zeta_{i-1}(c'), \zeta_i(c'))$, it holds that $w^*(c', \xi) = w_i(c')$. Also, it holds that $y_1(c') > \cdots > y_{k(c')}(c')$.

Lemma 3.3.1 provides the relationship between excess SPO risk and the optimal solution of (3.2) with respect to the difference $\Delta = \hat{c} - \bar{c}$ between the predicted cost vector $\hat{c}$ and the realized cost vector $\bar{c}$.

**Lemma 3.3.1.** *Let $\hat{c}, \bar{c} \in \mathbb{R}^d$ be given and define $\Delta := \hat{c} - \bar{c}$. Let $w_+ := w^*(\Delta)$ and $w_- := w^*(-\Delta)$, and let $y_+$ and $y_-$ denote the last elements of $w_+$ and $w_-$, respectively. If $\bar{c}^T(w^*(\hat{c}) - w^*(\bar{c})) \geq \epsilon$, then it holds that $\Delta^T(w_- - w_+) \geq \epsilon$. Additionally, if $\Delta = \kappa \cdot e_d$ for some $\kappa > 0$, then it holds that $(y_- - y_+)\kappa \geq \epsilon$.*

*Proof.* First we have $\hat{c}^T(w^*(\bar{c}) - w^*(\hat{c})) \geq 0$, and therefore it holds that $\Delta^T(w^*(\bar{c} + \Delta) - w^*(\bar{c})) \geq \bar{c}^T(w^*(\bar{c} + \Delta) - w^*(\bar{c})) \geq \epsilon$. Also, since $\Delta^T(w^*(\bar{c}) - w^*(\Delta)) \geq 0$ and $\Delta^T(w^*(-\Delta) - w^*(\bar{c} + \Delta)) \geq 0$, we have $\Delta^T(w_- - w_+) \geq \Delta^T(w^*(\bar{c} + \Delta) - w^*(\bar{c})) \geq \epsilon$. Moreover, when $\Delta = \kappa \cdot e_d$ for $\kappa > 0$, we have $\Delta^T w_- = \Delta^T w_1$ and $\Delta^T w_+ = \Delta^T w_k$, and therefore, it holds that $(y_- - y_+)\kappa \geq \epsilon$. □

Lemmas 3.3.2 and 3.3.3 provide two useful inequalities.

**Lemma 3.3.2.** *Suppose that $a_1, \ldots, a_n, b_1, \ldots, b_n \geq 0$ with $\sum_{i=1}^n a_i = \alpha$ and $\sum_{i=1}^n b_i = \beta$ for some $\alpha, \beta > 0$. Then for all $p \geq 0$, it holds that*

$$\sum_{i=1}^n b_i \left(1 + \frac{a_i^2}{b_i^2}\right)^{-p/2} \geq \frac{\beta}{(1 + \frac{\alpha}{\beta})^p}.$$

*Proof.* Let $\psi_i(a, b; p) = b_i(1 + \frac{a_i^2}{b_i^2})^{-p/2}$ and $\psi(a, b; p) = \sum_{i=1}^n \psi_i(a, b; p)$. For all $p \in \mathbb{R}$, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}p^2} \log(\psi(a, b; p)) = \frac{1}{4\psi^2(a, b; p)} \left( \sum_{i=1}^n \psi_i(a, b; p) \cdot \sum_{i=1}^n \psi_i(a, b; p) \log^2 \left(1 + \frac{a_i^2}{b_i^2}\right) - \left( \sum_{i=1}^n \psi_i(a, b; p) \log \left(1 + \frac{a_i^2}{b_i^2}\right) \right)^2 \right) \geq 0,$$

for $p \geq 0$. Therefore, for all $p \geq 0$ it holds that

$$\log \psi(a, b; p) \geq \log \psi(a, b; 0) + p \cdot (\log \psi(a, b; 0) - \log \psi(a, b; -1)).$$

Also, we have $\psi(a,b,0) = \beta$, and $\psi(a,b,-1) = \sum_{i=1}^{n}\sqrt{a_i^2+b_i^2} \le \sum_{i=1}^{n}(a_i+b_i) = \alpha+\beta$. Then, for all $p \ge 0$, it holds that $\psi(a,b;p) \ge \frac{\beta^{p+1}}{(\alpha+\beta)^p} = \frac{\beta}{(1+\frac{\alpha}{\beta})^p}$. □

**Lemma 3.3.3.** *Let $\hat{c}' \in \mathbb{R}^{d-1}$ be given with $\|\hat{c}'\|_2 = 1$, and let $\{w_i(\hat{c}')\}_{i=1}^k$, $\{y_i(\hat{c}')\}_{i=1}^k$, and $\{\zeta_i(\hat{c}')\}_{i=0}^k$ be the corresponding optimal solution sequence and phase transition location sequence as described in the definition and notation paragraph. Let $y_- = y_1(\hat{c}')$ and $y_+ = y_k(\hat{c}')$. Then it holds that*

$$\sum_{i=1}^{k-1}\left(1+3\zeta_i^2\right)^{-\frac{d-1}{2}}(y_i - y_{i+1}) \ge \Xi_{S,\hat{c}'}\cdot(y_- - y_+),$$

*where $\Xi_{S,\hat{c}'} = (1+\frac{2\sqrt{3}D_S}{y_- - y_+})^{1-d}$.*

*Proof.* Let $w_i'$ be the first $(d-1)$ element of $w_i$. Suppose $\zeta_{s-1} \le 0 < \zeta_s$ for some $s \in \{1,\dots,k\}$, then it holds that $\hat{c}'^T(w_i - w_{i+1}) = -\zeta_i(y_i - y_{i+1}) \ge 0$ for $i \in \{1,\dots,s-1\}$ and $\hat{c}'^T(w_i - w_{i+1}) = -\zeta_i(y_i - y_{i+1}) < 0$ for $i \in \{s,\dots,k-1\}$. Therefore, we know that

$$\sum_{i=1}^{k-1}\left|\hat{c}'^T(w_i - w_{i+1})\right| = \hat{c}'^T(w_1 + w_k - 2w_s) \le 2D_S.$$

Also, we have $\sum_{i=1}^{k-1}(y_i - y_{i+1}) = y_- - y_+$ and $|\zeta_i| = -\frac{|\hat{c}'^T(\hat{w}_i' - \hat{w}_{i+1}')|}{y_i - y_{i+1}}$. Therefore, by the result in Lemma 3.3.2, we have

$$\sum_{i=1}^{k-1}\left(1+3\zeta_i^2\right)^{-\frac{d-1}{2}}(y_i - y_{i+1}) \ge \frac{y_- - y_+}{(1+\frac{2\sqrt{3}D_S}{y_- - y_+})^{d-1}}.$$

□

Lemma 3.3.4 provide a lower bound of the conditional SPO+ risk condition on the first $(d-1)$ element of the realized cost vector.

**Lemma 3.3.4.** *Let $c' \in \mathbb{R}^{d-1}$ be a fixed vector and $\bar{\xi} \in \mathbb{R}$, $\sigma > 0$ be fixed scalars. Let a random variable $\xi$ satisfying $\mathbb{P}(\xi) \ge \alpha\cdot\mathcal{N}(\bar{\xi},\sigma^2)$ for all $\xi \in [-\sqrt{2D^2 - \|c'\|^2}, \sqrt{2D^2 - \|c'\|^2}]$. Let $c = (c',\xi) \in \mathbb{R}^d$, and sequence $\{w_i(c')\}_{i=0}^k$, $\{\zeta_i(c')\}_{i=0}^k$ defined as in the additional definitions and notation paragraph. Let $y_i$ denote the last element of vector $w_i$ for $i = 1,\dots,k$. Let $m_i = \sqrt{1+3\|\zeta_i(c')\|^2/\|c'\|^2}$ for $i = 1,\dots,k$. Suppose $\Delta = \kappa\cdot e_d$ for some $\kappa > 0$, then for all $\tilde{\kappa} \in [0,\kappa]$, it holds that*

$$\mathbb{E}_\xi\left[(c+2\Delta)^T(w^*(c) - w^*(c+2\Delta))\right] \ge \frac{\alpha\tilde{\kappa}\kappa e^{-\frac{3(\tilde{\kappa}^2+\bar{\xi}^2)}{2\sigma^2}}}{2}\cdot\sum_{i=1}^{k-1}\frac{e^{-\frac{3\zeta_i^2(c')}{2\sigma^2}}\mathbb{1}\{\|c'\| \le \frac{D}{m_i}\}}{\sqrt{2\pi}\sigma}(y_i - y_{i+1}).$$

*Proof.* Let $(w_1, \ldots, w_k) = \Omega(c')$ as defined in the additional definitions and notation paragraph, and suppose $w^*(c) = w_s$ and $w^*(c + 2\Delta) = w_t$ for some $s \leq t$. By the definition of $\{\xi_i(c')\}_0^k$, we know that $\xi \in [\zeta_{s-1}(c'), \zeta_s(c')]$ and $\xi + 2\kappa \in [\zeta_{t-1}(c'), \zeta_t(c')]$. Therefore, it holds that

$$
\begin{aligned}
&(c + 2\Delta)^T(w^*(c) - w^*(c + 2\Delta)) = (c + 2\Delta)^T(w_s - w_t) = \sum_{i=s}^{t-1}(c + 2\Delta)^T(w_i - w_{i+1}) \\
&= \sum_{i=s}^{t-1}(c + 2\Delta - (c', \zeta_i(c')))^T(w_i - w_{i+1}) = \sum_{i=s}^{t-1}(\xi + 2\kappa - \zeta_i(c')) \cdot e_d^T(w_i - w_{i+1}) \\
&= \sum_{i=1}^{k-1} \mathbb{1}\{\xi \in [\zeta_i - 2\kappa, \zeta_i]\} \cdot (\xi + 2\kappa - \zeta_i(c'))(y_i - y_{i+1}),
\end{aligned}
$$

where $y_i$ denotes the last element of $w_i$ for all $i = 1, \ldots, k$. When $\xi$ follows the normal distribution $\mathcal{N}(\bar{\xi}, \sigma^2)$, it holds that

$$
\begin{aligned}
&\mathbb{E}_\xi \left[\mathbb{1}\{\xi \in [\zeta_i - 2\kappa, \zeta_i]\} \cdot (\xi + 2\kappa - \zeta_i(c'))\right] \\
\geq &\mathbb{E}_\xi \left[\mathbb{1}\{\xi \in [\zeta_i - 2\tilde{\kappa}, \zeta_i]\} \cdot (\xi + 2\kappa - \zeta_i(c'))\right] \\
\geq &\int_{\zeta_i(c') - 2\tilde{\kappa}}^{\zeta_i(c')} \frac{\alpha e^{-\frac{(\xi - \bar{\xi})^2}{2\sigma^2}} \mathbb{1}\{\|c'\| \leq \frac{D}{m_i}\}}{\sqrt{2\pi\sigma^2}} \cdot (\xi + 2\kappa - \zeta_i(c'))\mathrm{d}\xi,
\end{aligned}
$$

for all $\tilde{\kappa} \in [0, \kappa]$. Therefore, it holds that

$$
\begin{aligned}
&\mathbb{E}_\xi \left[(c + 2\Delta)^T(w^*(c) - w^*(c + 2\Delta))\right] \\
\geq &\sum_{i=1}^{k-1}(y_i - y_{i+1})\tilde{\kappa}\kappa \cdot \frac{\alpha e^{-\frac{3(\zeta_i(c')^2 + \tilde{\kappa}^2 + \bar{\xi}^2)}{2\sigma^2}} \mathbb{1}\{\|c'\| \leq \frac{D}{m_i}\}}{2\sqrt{2\pi\sigma^2}} \\
= &\frac{\alpha\tilde{\kappa}\kappa e^{-\frac{3(\tilde{\kappa}^2 + \bar{\xi}^2)}{2\sigma^2}}}{2} \cdot \sum_{i=1}^{k-1} \frac{e^{-\frac{3\zeta_i^2(c')}{2\sigma^2}} \mathbb{1}\{\|c'\| \leq \frac{D}{m_i}\}}{\sqrt{2\pi\sigma^2}}(y_i - y_{i+1}).
\end{aligned}
$$

□

Lemma 3.3.5 provide a lower bound of the conditional SPO+ risk when the distribution of $c = (c', \epsilon)$ is well behaved.

**Lemma 3.3.5.** *Let $\bar{c}' \in \mathbb{R}^{d-1}$ be a fixed vector and $\bar{\xi} \in \mathbb{R}$, $\sigma > 0$ be fixed scalars. Let $c' \in \mathbb{R}^{d-1}$ be a random vector satisfying $\mathbb{P}(c') \geq \mathcal{N}(\bar{c}', \sigma^2 I_{d-1})$ for all $\|c'\|_2^2 \leq 2D^2$, and let $\xi \in \mathbb{R}$ be a random variable satisfying $\mathbb{P}(\xi|c') \geq \alpha \cdot \mathcal{N}(\bar{\xi}, \sigma^2)$ for all $\xi \in [-\sqrt{2D^2 - \|c'\|^2}, \sqrt{2D^2 - \|c'\|^2}]$. Define $\Xi_S := (1 + \frac{2\sqrt{3}D_S}{d_S})^{1-d}$. Suppose $\Delta = \kappa \cdot e_d$ for some $\kappa > 0$, then for all $\tilde{\kappa} \in [0, \kappa]$, it holds that*

$$
\mathbb{E}_{c',\xi}\left[(c + 2\Delta)^T(w^*(c) - w^*(c + 2\Delta))\right] \geq \frac{\alpha\tilde{\kappa}\kappa e^{-\frac{3\tilde{\kappa}^2 + 3\bar{\xi}^2 + \|\bar{c}'\|_2^2}{2\sigma^2}}}{4\sqrt{2\pi\sigma^2}} \cdot \frac{\gamma(\frac{d-1}{2}, D^2)}{\Gamma(\frac{d-1}{2})} \cdot \Xi_S(y_- - y_+).
$$

*Proof.* By result in Lemma 3.3.4, it holds that

$$\begin{aligned}
&\mathbb{E}_{c',\xi}\left[(c+2\Delta)^T(w^*(c)-w^*(c+2\Delta))\right] \\
&\geq \frac{\alpha\tilde{\kappa}\kappa e^{-\frac{3(\tilde{\kappa}^2+\tilde{\xi}^2)}{2\sigma^2}}}{2}\cdot\mathbb{E}_{c'}\left[\sum_{i=1}^{k(c')-1}\frac{e^{-\frac{3\zeta_i^2(c')}{2\sigma^2}}\mathbb{1}\{\|c'\|\leq\frac{D}{m_i}\}}{\sqrt{2\pi\sigma^2}}(y_i(c')-y_{i+1}(c'))\right].
\end{aligned}$$

For any $c'\in\mathbb{R}^{d-1}$, let $r=\|c'\|_2$ and $\hat{c}'=\frac{c'}{r}$. We know that $k(c')=k(\hat{c}')$, $\zeta_i(c')=r\zeta_i(\hat{c}')$, $w_i(c')=w_i(\hat{c}')$, and $y_i(c')=y_i(\hat{c}')$. Then we have

$$\begin{aligned}
&\mathbb{E}_{c'}\left[\sum_{i=1}^{k(c')-1}\frac{e^{-\frac{3\zeta_i^2(c')}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}(y_i(c')-y_{i+1}(c'))\right] \\
&=\int_{\mathbb{S}^{d-2}}\int_0^\infty\sum_{i=1}^{k(\hat{c}')-1}\frac{e^{-\frac{3r^2\zeta_i^2(\hat{c}')}{2\sigma^2}\mathbb{1}\{r\leq\frac{D}{m_i}\}}}{\sqrt{2\pi\sigma^2}}(y_i(\hat{c}')-y_{i+1}(\hat{c}'))r^{d-2}\mathbb{P}_{c'}(r\hat{c}')\mathrm{d}r\mathrm{d}\hat{c}',
\end{aligned}$$

where $\mathbb{S}^{d-2}=\{\hat{c}'\in\mathbb{R}^{d-1}:\|\hat{c}'\|_2=1$. For fixed $\hat{c}'\in\mathbb{S}^{d-2}$ with $\hat{c}'^T\bar{c}'\geq 0$ and $i\in\{1,\ldots,k(\hat{c}')-1\}$, we have

$$\begin{aligned}
\int_0^{\frac{D}{m_i}}\frac{e^{-\frac{3r^2\zeta_i^2(\hat{c}')}{2\sigma^2}}}{\sqrt{2\pi\sigma}}r^{d-2}\mathbb{P}_{c'}(r\hat{c}')\mathrm{d}r &= \int_0^{\frac{D}{m_i}}\frac{e^{-\frac{3r^2\zeta_i^2(\hat{c}')}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\cdot\frac{e^{-\frac{\|r\hat{c}'-\bar{c}'\|_2^2}{2\sigma^2}}}{(2\pi\sigma^2)^{\frac{d-1}{2}}}\cdot r^{d-2}\mathrm{d}r \\
&\geq \int_0^{\frac{D}{m_i}}\frac{e^{-\frac{3r^2\zeta_i^2(\hat{c}')}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\cdot\frac{e^{-\frac{r^2+\|\bar{c}'\|_2^2}{2\sigma^2}}}{(2\pi\sigma^2)^{\frac{d-1}{2}}}\cdot r^{d-2}\mathrm{d}r \\
&= \frac{e^{-\frac{\|\bar{c}'\|_2^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\cdot\frac{\gamma(\frac{d-1}{2},\frac{D^2(1+3\zeta_i^2(\hat{c}'))}{m_i^2})}{2\pi^{\frac{d-1}{2}}}\cdot(1+3\zeta_i^2(\hat{c}'))^{-\frac{d-1}{2}} \\
&\geq \frac{e^{-\frac{\|\bar{c}'\|_2^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\cdot\frac{\gamma(\frac{d-1}{2},D^2)}{2\pi^{\frac{d-1}{2}}}\cdot(1+3\zeta_i^2(\hat{c}'))^{-\frac{d-1}{2}},
\end{aligned}$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete Gamma function. By Lemma 3.3.3, it holds that

$$\sum_{i=1}^{k(\hat{c}')-1}(1+3\zeta_i^2(\hat{c}'))^{-\frac{d-1}{2}}(y_i(\hat{c}')-y_{i+1}(\hat{c}'))\geq\Xi_{S,\hat{c}'}\cdot(y_--y_+). \tag{3.6}$$

Therefore, it holds that

$$\begin{aligned}
&\mathbb{E}_{c'}\left[\sum_{i=1}^{k(c')-1}\frac{e^{-\frac{3\zeta_i^2(c')}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}(y_i(c')-y_{i+1}(c'))\right]\\
\geq&\int_{\mathbb{S}^{d-2}}\mathbb{1}\{\hat{c}'^T\bar{c}'\geq 0\}\cdot\frac{e^{-\frac{\|\bar{c}'\|_2^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\cdot\frac{\gamma(\frac{d-1}{2},D^2)}{2\pi^{\frac{d-1}{2}}}\cdot\Xi_{S,c'}(y_--y_+)\mathrm{d}\hat{c}'\\
\geq&\frac{e^{-\frac{\|\bar{c}'\|_2^2}{2\sigma^2}}}{2\sqrt{2\pi\sigma^2}}\cdot\frac{\gamma(\frac{d-1}{2},D^2)}{\Gamma(\frac{d-1}{2})}\cdot\Xi_S(y_--y_+),
\end{aligned}$$

and finally we get

$$\begin{aligned}
&\mathbb{E}_{c',\xi}\left[(c+2\Delta)^T(w^*(c)-w^*(c+2\Delta))\right]\\
\geq&\frac{\alpha\tilde{\kappa}\kappa e^{-\frac{3(\tilde{\kappa}^2+\bar{\xi}^2)}{2\sigma^2}}}{2}\cdot\frac{e^{-\frac{\|\bar{c}'\|_2^2}{2\sigma^2}}}{2\sqrt{2\pi\sigma^2}}\cdot\frac{\gamma(\frac{d-1}{2},D^2)}{\Gamma(\frac{d-1}{2})}\cdot\Xi_S(y_--y_+)\\
=&\frac{\alpha\tilde{\kappa}\kappa e^{-\frac{3\tilde{\kappa}^2+3\bar{\xi}^2+\|\bar{c}'\|_2^2}{2\sigma^2}}}{4\sqrt{2\pi\sigma^2}}\cdot\frac{\gamma(\frac{d-1}{2},D^2)}{\Gamma(\frac{d-1}{2})}\cdot\Xi_S(y_--y_+).
\end{aligned}$$

□

Now we provide the proof of Theorem 3.3.1.

*Proof of Theorem 3.3.1.* Without loss of generality, we assume $d_S > 0$. Otherwise, the constant $\Xi_S$ will be zero and (3.4) will be a trivial bound. Let $\kappa = \|\Delta\|_2$ and $A \in \mathbb{R}^{d\times d}$ be an orthogonal matrix such that $A^T\Delta = \kappa\cdot e_d$ for $e_d = (0,\ldots,0,1)^T$. We implement a change of basis and let the new basis be $A = (a_1,\ldots,a_d)$. With a slight abuse of notation, we keep the notation the same after the change of basis, for example, now the vector $\Delta$ equals to $\kappa\cdot e_d$. Since the excess SPO risk of $\hat{c} = \bar{c}+\Delta$ is at least $\epsilon$, we have $\kappa(y_- - y_+)\geq\epsilon$. Let $\tilde{\kappa} = \min\{\kappa,\sigma\}$. Then it holds that $\tilde{\kappa}\exp(-\frac{3\tilde{\kappa}^2}{2\sigma^2}) \geq \min\{\kappa,\sigma\}\cdot\exp(-\frac{3}{2})$. By Lemma 3.3.5, we know that

$$\begin{aligned}
R_{\mathrm{SPO}}(\hat{c}) &= \mathbb{E}_c\left[(c+2\Delta)^T(w^*(c)-w^*(c+2\Delta))\right]\\
&\geq\frac{\tilde{\kappa}\kappa e^{-\frac{3\tilde{\kappa}^2+3\bar{\xi}^2+\|\bar{c}'\|_2^2}{2\sigma^2}}\gamma(\frac{d-1}{2},D^2)}{4\sqrt{2\pi\sigma^2}\Gamma(\frac{d-1}{2})}\cdot\Xi_S(y_--y_+),
\end{aligned}$$

where $\bar{c}'$ is the first $(d-1)$ elements of $\bar{c}$, $\bar{\xi}$ is the last element of $\bar{c}$, and $3\bar{\xi}^2+\|\bar{c}'\|_2^2 \leq 3\|\bar{c}\|_2^2 = 3\alpha^2\sigma^2$. Then we can conclude that

$$R_{\mathrm{SPO+}}(\hat{c}) - R^*_{\mathrm{SPO+}} \geq \frac{\alpha\Xi_S\gamma(\frac{d-1}{2},D^2)\cdot\epsilon}{4\sqrt{2\pi}e^{\frac{3(1+\beta^2)}{2}}\Gamma(\frac{d-1}{2})}\cdot\min\left\{\frac{\kappa}{\sigma},1\right\}.$$

Furthermore, since $\frac{\kappa}{\sigma} \geq \frac{\kappa}{M} \geq \frac{\epsilon}{(y_- - y_+)M} \geq \frac{\epsilon}{D_S M}$, we have

$$R_{\text{SPO+}}(\hat{c}) - R^*_{\text{SPO+}} \geq \frac{\alpha \Xi_S \gamma(\frac{d-1}{2}, D^2)}{4\sqrt{2\pi} e^{\frac{3(1+\beta^2)}{2}} \Gamma(\frac{d-1}{2})} \cdot \min\left\{ \frac{\epsilon^2}{D_S M}, \epsilon \right\}.$$

□

### 3.3.1 Faster Rates with Low Near-Degeneracy Condition

In this section, we develop the *fast rates* in the SPO+ risk bounds when the feasible region $S$ is a polyhedron. We improve the sample complexity from $\mathcal{O}(n^{-\frac{1}{4}})$ to $\mathcal{O}(n^{-\frac{\kappa+1}{2\kappa+4}})$ where $\kappa \in (0, \infty)$ is the low near-degeneracy parameter. The main assumption is equivalent to the *low-noise condition* defined in [41], and the assumption on the *near-degeneracy* defined in [28], as we will show later.

Let $S^\angle$ be the set of all extreme points of the polyhedral feasible set $S$. For all feature vector $x \in \mathcal{X}$, let $g^*(x) := \mathbb{E}[c|x]$ denote the conditional expectation, and let $W^*(x)$ be the set of optimal solutions, namely $W^*(x) := \arg\min_{w \in S^\angle} g^*(x)^T w$. Define

$$A(x) := \begin{cases} \min\limits_{w \in S^\angle \setminus W^*(x)} \left\{ g^*(x)^T w \right\} - g^*(x)^T w^*(g^*(x)), & S^\angle \neq W^*(x), \\ 0, & S^\angle = W^*(x). \end{cases}$$

**Assumption 3.3.1.** *There exists some constants $\kappa, \gamma > 0$ such that*

$$\mathbb{P}\left(A(X) \in (0, \delta)\right) \leq \gamma \cdot \delta^\kappa, \quad \forall \delta > 0. \tag{3.7}$$

The $A(\cdot)$ measures the sub-optimality of the second-best choice with feature vector $x$. We notice that the $A(\cdot)$ is at the same order of the margin defined as the distance to degeneracy.

**Lemma 3.3.6.** *Let $x \in \mathcal{X}$ and $\bar{c} = \mathbb{E}[c|x]$. Suppose $W^*(\bar{c})$ is a singleton. Let $\mathcal{C}^\circ$ denote the set of degenerate cost vectors and $v_S(\bar{c}) := \inf_{c \in \mathcal{C}^\circ} \{\|c - \bar{c}\|_*\}$ denote the distance to degeneracy of $\bar{c}$. Then it holds that*

$$\frac{A(x)}{D_S} \leq v_S(\bar{c}) \leq \frac{A(x)}{m_S},$$

*where $m_S = \min_{w, w' \in S^\angle, w \neq w'} \|w - w'\|$.*

Therefore, when $\mathbb{P}(\bar{c}(x) \in \mathcal{C}^\circ) = 0$, then Assumption 3.3.1 is equivalent to the following Assumption 3.3.2 on the distance to degeneracy the same $\kappa$.

**Assumption 3.3.2.** *There exists some constants $\kappa, \gamma > 0$ such that*

$$\mathbb{P}\left(v_S(\bar{c}(x)) \in [0, \delta)\right) \leq \gamma \cdot \delta^\kappa, \quad \forall \delta > 0. \tag{3.8}$$

Given the low-noise condition, we can extend Theorem 3 in [17] for binary classification problem to the general SPO setting.

**Theorem 3.3.2.** *Suppose Assumption 3.3.1 holds. Let $\ell(\cdot)$ be a surrogate loss function, and $\psi(\cdot)$ be its calibration function, then there exists some constant $C > 0$ such that*

$$C\left(R(g) - R^*\right)^{\kappa/(\kappa+1)} \cdot \psi\left(\frac{\left(R(g) - R^*\right)^{1/(\kappa+1)}}{2C}\right) \leq R_\ell(g) - R_\ell^*. \tag{3.9}$$

In polyhedral feasible region case, an upper bound on the original calibration function for SPO+ loss is $\psi(\epsilon) \geq C' \cdot \epsilon^2$. Therefore, by applying the results into Theorem 3.3.2, we have the following guarantee on the SPO+ risk bounds and sample complexity.

**Proposition 3.3.1.** *Suppose Assumption 3.3.1 holds, it holds that*

$$C'' \cdot \left(R(g) - R^*\right)^{(\kappa+2)/(\kappa+1)} \leq R_{\text{SPO+}}(g) - R_{\text{SPO+}}^*. \tag{3.10}$$

*Also, the sample complexity is $R(\hat{g}^n) - R^* \leq \mathcal{O}(n^{-\frac{\kappa+1}{2\kappa+4}})$ when $\mathfrak{R}^n(\mathcal{H}) \leq \mathcal{O}(n^{-\frac{1}{2}})$.*

Now we provide the proofs of Lemma 3.3.6 and Theorem 3.3.2.

*Proof of Lemma 3.3.6.* Since $W^*(\bar{c})$ is a singleton, let $w^* = w^*(\bar{c})$. Suppose $\hat{c} \in \mathcal{C}$ has multiple optimality with $\hat{c}^T w' = \hat{c}^T w'' \leq \hat{c}^T w^*$. By the definition of $A(x)$, it holds that $\bar{c}^T w' - \bar{c}^T w^* \geq A(x)$. Therefore, we have $A(x) \leq (\hat{c} - \bar{c})^T (w^* - w') \leq \|\hat{c} - \bar{c}\|_* \cdot \|w^* - w'\|$ and hence $\|\hat{c} - \bar{c}\|_* \geq \frac{A(x)}{\|w^* - w'\|} \geq \frac{A(x)}{D_S}$. Taking the infimum over $\hat{c}$ we arrive at $v_S(\bar{c}) \geq \frac{A(x)}{D_S}$.

On the other hand, pick $w' = \arg\min_{w \in S^\angle \setminus \{w^*\}} \frac{\bar{c}^T(w - w^*)}{\|w - w^*\|}$. Let $\tilde{c} = \bar{c} - \frac{\bar{c}^T(w' - w^*)}{\|w' - w^*\|} \cdot u$, where $u \in \mathbb{R}^d$ satisfies $\|u\|_* = 1$ and $u^T(w' - w^*) = \|w' - w^*\|$. For any $w'' \in S^\angle \setminus \{w^*, w'\}$, it holds that

$$\begin{aligned}
\tilde{c}^T w'' - \tilde{c}^T w^* &= \bar{c}^T(w'' - w^*) - \frac{\bar{c}^T(w' - w^*)}{\|w' - w^*\|} \cdot u^T(w' - w^*) \\
&\geq \bar{c}^T(w'' - w^*) - \frac{\bar{c}^T(w' - w^*)}{\|w' - w^*\|} \cdot \|w'' - w^*\| \geq 0.
\end{aligned}$$

Therefore, we have $\tilde{c} \in \mathcal{C}^\circ$ and hence

$$v_S(\bar{c}) \leq \|\tilde{c} - \bar{c}\| = \min_{w \in S^\angle \setminus \{w^*\}} \frac{\bar{c}^T(w - w^*)}{\|w - w^*\|} \leq \min_{w \in S^\angle \setminus \{w^*\}} \frac{\bar{c}^T(w - w^*)}{m_S} = \frac{v_S(\bar{c})}{m_S}.$$

□

First we extend the Lemma 5 in [17] to the general SPO setting. Let $w_g^*(x) = w^*(g(x))$ for any $x \in \mathcal{X}$ and $\mathcal{T}_g = \{x \in \mathcal{X} : g^*(x)^T(w_g^*(x) - w_{g^*}^*(x)) > 0\}$.

**Lemma 3.3.7** (Lemma 5 in [17]). *Suppose Assumption 3.3.1 holds. For any prediction function $g(\cdot)$ and feature vector $x \in \mathcal{X}$, define $w_g^*(x) = w^*(g(x))$, and define $\mathcal{T}_g = \{x \in \mathcal{X} : g^*(x)^T(w_g^*(x) - w_{g^*}^*(x)) > 0\}$. Then there exists some constant $\gamma' > 0$, such that*

$$R_{\text{SPO}}(g) - R_{\text{SPO}}^* \geq \gamma' \cdot \mathbb{P}(X \in \mathcal{T}_g)^{(\kappa+1)/\kappa}.$$

*Proof.* For any $\epsilon > 0$, it holds that

$$\begin{aligned} R_{\text{SPO}}(g) - R_{\text{SPO}}^* &= \mathbb{E}\left[g^*(X)^T(w_g^*(X) - w_{g^*}^*(X))\right] \\ &\geq \epsilon \cdot \mathbb{P}\left(g^*(X)^T(w_g^*(X) - w_{g^*}^*(X)) \geq \epsilon\right) \\ &\geq \epsilon \cdot \left(\mathbb{P}(X \in \mathcal{T}_g) - \mathbb{P}\left(g^*(X)^T(w_g^*(X) - w_{g^*}^*(X)) \in (0, \epsilon)\right)\right) \\ &\geq \epsilon \cdot \left(\mathbb{P}(X \in \mathcal{T}_g) - \gamma \cdot \epsilon^\kappa\right), \end{aligned}$$

where the last inequality comes from Assumption 3.3.1. By setting $\epsilon \leftarrow \left(\mathbb{P}(X \in \mathcal{T}_g)/\gamma(1+\kappa)\right)^{1/\kappa}$, we have

$$R_{\text{SPO}}(g) - R_{\text{SPO}}^* \geq \gamma' \cdot \mathbb{P}(X \in \mathcal{T}_g)^{(\kappa+1)/\kappa}.$$

□

*Proof of Theorem 3.3.2.* For any prediction function $g(\cdot) : \mathcal{X} \to \mathbb{R}^d$, let $d_g(x) := g^*(x)^T(w_g^*(x) - w_{g^*}^*(x))$ denote the optimality gap at feature vector $x \in \mathcal{X}$. Then, for any $\epsilon > 0$, it holds that

$$\mathbb{E}_X[d_g(X)] = \mathbb{E}_X[d_g(X)\mathbb{1}\{d_g(X) \in (0, \epsilon)\}] + \mathbb{E}_X[d_g(X)\mathbb{1}\{d_g(X) > \epsilon\}].$$

For the first term, we have

$$\begin{aligned} \mathbb{E}_X[d_g(X)\mathbb{1}\{d_g(X) \in (0, \epsilon)\}] &\leq \epsilon \cdot \mathbb{E}_X[\mathbb{1}\{d_g(X) \in (0, \epsilon)\}] \\ &\leq \epsilon \cdot \mathbb{E}_X[\mathbb{1}\{d_g(X) > 0\}] \\ &\leq \epsilon \cdot \left(\frac{R_{\text{SPO}}(g) - R_{\text{SPO}}^*}{\gamma'}\right)^{\kappa/(\kappa+1)}, \end{aligned}$$

where the last inequality comes from Lemma 3.3.7. For the second term, it holds that for any $x : d_g(x) \in [0, \epsilon]$ we have

$$d_g(x)\mathbb{1}\{d_g(x) > \epsilon\} = 0 \leq \frac{\epsilon}{\psi(\epsilon)} \cdot \psi(d_g(x)),$$

and for any $x : d_g(x) > \epsilon$ we have

$$d_g(x)\mathbb{1}\{d_g(x) > \epsilon\} = d_g(x) \leq \frac{\epsilon}{\psi(\epsilon)} \cdot \psi(d_g(x)),$$

where the last inequality holds since $\psi(\cdot)$ is convex and increasing, and $\psi(0) = 0$. Therefore, it holds that

$$\mathbb{E}_X[d_g(X)\mathbb{1}\{d_g(X) > \epsilon\}] \leq \mathbb{E}_X\left[\frac{\epsilon}{\psi(\epsilon)} \cdot \psi(d_g(X))\right] \leq \frac{\epsilon}{\psi(\epsilon)} \cdot (R_\ell(g) - R_\ell^*),$$

where the second inequality holds since

$$\psi(d_g(x)) = \psi(\mathbb{E}_{c|x}[\ell_{\text{SPO}}(g(x), c) - \ell_{\text{SPO}}(g^*(x), c)]) \leq \mathbb{E}_{c|x}[\ell(g(x), c) - \ell(g^*(x), c)].$$

Therefore, we have

$$\begin{aligned} R_{\text{SPO}}(g) - R^*_{\text{SPO}} = \mathbb{E}_X[d_g(X)] &= \mathbb{E}_X[d_g(X)\mathbb{1}\{d_g(X) \in (0, \epsilon)\}] + \mathbb{E}_X[d_g(X)\mathbb{1}\{d_g(X) > \epsilon\}] \\ &\leq \epsilon \cdot \left(\frac{R_{\text{SPO}}(g) - R^*_{\text{SPO}}}{\gamma'}\right)^{\kappa/(\kappa+1)} + \frac{\epsilon}{\psi(\epsilon)} \cdot (R_\ell(g) - R^*_\ell). \end{aligned}$$

By choosing $\epsilon \leftarrow \frac{\gamma'}{2} \cdot \left(\frac{R_{\text{SPO}}(g) - R^*_{\text{SPO}}}{\gamma'}\right)^{1/(\kappa+1)}$, we have

$$\frac{R_{\text{SPO}}(g) - R^*_{\text{SPO}}}{2} \leq \frac{\frac{\gamma'}{2} \cdot \left(\frac{R_{\text{SPO}}(g) - R^*_{\text{SPO}}}{\gamma'}\right)^{1/(\kappa+1)}}{\psi\left(\frac{\gamma'}{2} \cdot \left(\frac{R_{\text{SPO}}(g) - R^*_{\text{SPO}}}{\gamma'}\right)^{1/(\kappa+1)}\right)} \cdot (R_\ell(g) - R^*_\ell),$$

which leads to

$$C\left(R(g) - R^*\right)^{\kappa/(\kappa+1)} \cdot \psi\left(\frac{\left(R(g) - R^*\right)^{1/(\kappa+1)}}{2C}\right) \leq R_\ell(g) - R^*_\ell.$$

□

## 3.4 Risk Bounds and Calibration for Strongly Convex Level Sets

In this section, we develop improved risk bounds for the SPO+ loss function under the assumption that the feasible region is the level set of a strongly convex and smooth function, formalized in Assumption 3.4.1 below. The assumption allows the domain of the strongly convex function to be a subset of $\mathbb{R}^d$. In particular, we define the domain set $T \subseteq \mathbb{R}^d$ by $T := \{w \in \mathbb{R}^d : h_i^T w = s_i \ \forall i \in [m_1], t_j(w) < 0 \ \forall j \in [m_2]\}$, where $h_i \in \mathbb{R}^d$ and $s_i \in \mathbb{R}$ for $i \in [m_1]$, and $t_j(\cdot) : \mathbb{R}^d \to \mathbb{R}$ are convex functions for $j \in [m_2]$. Clearly, when $m_1 = m_2 = 0$, the set $T$ is the entire vector space $\mathbb{R}^d$. Also, let the closure of $T$ be $\bar{T} = \{w \in \mathbb{R}^d : h_i^T w = s_i \ \forall i \in [m_1], t_j(w) \leq 0 \ \forall j \in [m_2]\}$, and with a slight abuse of notation, let the (relative) boundary of $T$ be $\partial T := \bar{T} \backslash T$. For any function defined on $T$, we consider the (relative) lower limit be $\underline{\lim}_{w \to \partial T} = \inf_{\delta > 0} \sup_{w \in T : d(w, \partial T) \leq \delta} f(w)$, where the distance function $d(\cdot, \cdot)$ is defined as $d(w, \partial T) = \min_{w' \in \partial T} \|w - w'\|_2$.

**Assumption 3.4.1.** *For a given norm $\|\cdot\|$, let $f : T \to \mathbb{R}$ be a $\mu$-strongly convex function on $T$ for some $\mu > 0$. Assume that the feasible region $S$ is defined by $S = \{w \in T : f(w) \leq r\}$ for some constant $r$ satisfying $\underline{\lim}_{w \to \partial T} f(w) > r > f_{\min} := \min_{w \in T} f(w)$. Additionally, assume that $f$ is $L$-smooth on $S$ for some $L \geq \mu$.*

The assumption allows for a broad choice of feasible regions, for instance, any bounded $\ell_q$ ball for any $q \in (1, 2]$ and the probability simplex with entropy constraint. The latter example, which can also be thought of as portfolio allocation with an entropy constraint, is considered in the experiments in Section 3.5.

As in the polyhedral case, the distribution class $\mathcal{P}_{\text{cont, symm}}$ is not restrictive enough to derive a meaningful lower bound on the calibration function of the SPO+ loss. We instead consider two related classes of rotationally symmetric distributions with bounded conditional coefficient of variation. These distribution classes are formally defined in Definition 3.4.1 below, and include the multi-variate Gaussian, Laplace, and Cauchy distributions as special cases.

**Definition 3.4.1.** *Let $A$ be a given positive definite matrix. We define $\mathcal{P}_{\text{rot symm},A}$ as the class of distributions with conditional rotational symmetry in the norm $\|\cdot\|_{A^{-1}}$, namely*

$$\mathcal{P}_{\text{rot symm},A} := \{\mathbb{P} : \forall x \in \mathcal{X}, \exists q(\cdot) : [0,\infty] \to [0,\infty] \text{ such that } \mathbb{P}(c|x) = q(\|c - \bar{c}\|_{A^{-1}})\}.$$

*Let $\bar{c}$ denote the conditional expectation $\bar{c} = \mathbb{E}[c|x]$. For constants $\alpha \in (0, 1]$ and $\beta > 0$, define*

$$\mathcal{P}_{\beta,A} := \left\{\mathbb{P} \in \mathcal{P}_{\text{rot symm},A} : \mathbb{E}_{c|x}[\|c - \bar{c}\|^2_{A^{-1}}] \le \beta^2 \cdot \|\bar{c}\|^2_{A^{-1}},\ \forall x \in \mathcal{X}\right\},$$

*and*

$$\mathcal{P}_{\alpha,\beta,A} := \left\{\mathbb{P} \in \mathcal{P}_{\text{rot symm},A} : \mathbb{P}_{c|x}(\|c - \bar{c}\|_{A^{-1}} \le \beta \cdot \|\bar{c}\|_{A^{-1}}) \ge \alpha,\ \forall x \in \mathcal{X}\right\}.$$

Under the above assumptions, Theorem 3.4.1 demonstrates that the calibration function of the SPO+ loss is $\mathcal{O}(\epsilon)$, significantly strengthening our result in the polyhedral case. The results hold for the aforementioned case where the domain of $f(\cdot)$ may be a subset of $\mathbb{R}^d$, which includes the entropy constrained portfolio allocation problem for example. Also, the results provide lower bounds of the calibration function of two different distribution classes, which include the multi-variate Gaussian, Laplace, and Cauchy distribution.

**Theorem 3.4.1.** *Suppose that Assumption 3.4.1 holds with respect to the norm $\|\cdot\|_A$ for some positive definite matrix $A$. Then, for any $\epsilon > 0$, it holds that $\hat{\delta}_{\ell_{\text{SPO+}}}(\epsilon; \mathcal{P}_{\beta,A}) \ge \frac{\mu^{9/2}}{4(1+\beta^2)L^{9/2}} \cdot \epsilon$ and $\hat{\delta}_{\ell_{\text{SPO+}}}(\epsilon; \mathcal{P}_{\alpha,\beta,A}) \ge \frac{\alpha\mu^{9/2}}{4(1+\beta^2)L^{9/2}} \cdot \epsilon$.*

Let us now provide some more intuition on the parameters involved in the definitions of the distribution classes $\mathcal{P}_{\beta,A}$ and $\mathcal{P}_{\alpha,\beta,A}$ and their roles in Theorem 3.4.1. In both cases, the parameter $\beta$ controls the concentration of the distribution of cost vector around the conditional mean. The more concentrated the distribution is, the better the bounds in Theorem 3.4.1 are. In the case of $\mathcal{P}_{\alpha,\beta,A}$, the parameter $\alpha$ relates to the probability that the cost vector is "relatively close" to the conditional mean. When $\alpha$ is larger, the cost vector is more likely to be close to the conditional mean and the bound will be better.

Our analysis for the calibration function (the proof of Theorem 3.4.1) relies on the following two lemmas, which utilize the special structure of the feasible region to strengthen the "first-order optimality" and provide a "Lipschitz-like" continuity of the optimization oracle

$w^*(\cdot)$. We want to mention that some of the results in the following two lemmas generalize the results in [28] to the cases where the feasible region $S$ is defined on a subspace of $\mathbb{R}^d$ rather than an open set in $\mathbb{R}^d$. Let $H$ denote the linear subspace defined by the linear combination of all $h_j$, namely $H = \text{span}(\{h_j\}_{j=1}^{m_2})$, and let $H^\perp$ denote its orthogonal complement, namely $H^\perp = \{w \in \mathbb{R}^d : h_j^T w = 0, \forall j \in [m_2]\}$. Also, for any $c \in \mathbb{R}^d$, let $\text{proj}_{H^\perp}(c)$ denote its projection onto $H^\perp$. A special instance is when $m_1 = m_2 = 0$, and thus $H^\perp = \mathbb{R}^d$. The first such lemma strengthens the optimality guarantees of (3.2) and provides both upper and lower bounds of the SPO loss.

**Lemma 3.4.1.** *Suppose Assumption 3.4.1 holds with respect to a generic norm $\|\cdot\|$. Then, for any $c_1, c_2 \in \mathbb{R}^d$, it holds that*

$$c_1^T(w - w^*(c_1)) \geq \frac{\mu}{2\sqrt{2L(r - f_{\min})}} \|\text{proj}_{H^\perp}(c_1)\|_* \|w - w^*(c_1)\|^2, \quad \forall w \in S, \tag{3.11}$$

*and*

$$c_1^T(w^*(c_2) - w^*(c_1)) \leq \frac{L}{2\sqrt{2\mu(r - f_{\min})}} \|\text{proj}_{H^\perp}(c_1)\|_* \|w^*(c_1) - w^*(c_2)\|^2. \tag{3.12}$$

The two constants are the same since Theorem 12 in [43] showed that set $S$ is a $\frac{\mu}{\sqrt{2Lr}}$-strongly convex set. The following lemma provides upper and lower bounds in the difference between two optimal decision based on the difference between the two normalized cost vector.

**Lemma 3.4.2.** *Suppose that Assumption 3.4.1 holds with respect to a generic norm $\|\cdot\|$. Let $c_1, c_2 \in \mathbb{R}^d$ be such that $\text{proj}_{H^\perp}(c_1), \text{proj}_{H^\perp}(c_2) \neq 0$, then it holds that*

$$\|w^*(c_1) - w^*(c_2)\| \geq \frac{\sqrt{2\mu(r - f_{\min})}}{L} \cdot \left\| \frac{\text{proj}_{H^\perp}(c_1)}{\|\text{proj}_{H^\perp}(c_1)\|_*} - \frac{\text{proj}_{H^\perp}(c_2)}{\|\text{proj}_{H^\perp}(c_2)\|_*} \right\|_*,$$

*and*

$$\|w^*(c_1) - w^*(c_2)\| \leq \frac{\sqrt{2L(r - f_{\min})}}{\mu} \cdot \left\| \frac{\text{proj}_{H^\perp}(c_1)}{\|\text{proj}_{H^\perp}(c_1)\|_*} - \frac{\text{proj}_{H^\perp}(c_2)}{\|\text{proj}_{H^\perp}(c_2)\|_*} \right\|_*.$$

Note that the lower bound of $c_1^T(w - w^*(c_1))$ in Lemma 3.4.1 and the upper bound of $\|w^*(c_1) - w^*(c_2)\|$ in Lemma 3.4.2 match bounds developed by [28]. Indeed, although [28] study the more general case of strongly convex sets, the constants are the same since Theorem 12 of [43] demonstrates that our set $S$ is a $\frac{\mu}{\sqrt{2L(r-f_{\min})}}$-strongly convex set. However, the upper bounds in Lemmas 3.4.1 and 3.4.2 appear to be novel and rely on the special properties of strongly convex *level* sets. It is important to emphasize that we generally do not expect all of the bounds in Lemmas 3.4.1 and 3.4.2 to holds for polyhedral sets. Indeed, for a polyhedron the optimization oracle $w^*(\cdot)$ is generally discontinuous at cost vectors that have multiple optimal solutions. The properties in Lemmas 3.4.1 and 3.4.2 drive the proof of Theorem 3.4.1 and hence lead to the improvement from $\mathcal{O}(\epsilon^2)$ in the polyhedral case to $\mathcal{O}(\epsilon)$ in the strongly convex level set case.

By following similar arguments as in the derivation of Corollaries 3.3.1 and 3.4.1 presents the sample complexity of the SPO+ method when the feasible region is a strongly convex level set.

**Corollary 3.4.1.** *Suppose that Assumption 3.4.1 holds with respect to the norm $\|\cdot\|_A$ for some positive definite matrix $A$. Suppose further that the optimal predictor $g^*(x) = \mathbb{E}[c|x]$ is in the hypothesis class $\mathcal{H}$, and there exists a constant $C'$ such that $\mathfrak{R}^n(\mathcal{H}) \leq \frac{C'}{\sqrt{n}}$. Let $\hat{g}^n_{\text{SPO+}}$ denote the predictor which minimizes the empirical SPO+ risk $\hat{R}^n_{\text{SPO+}}(\cdot)$ over $\mathcal{H}$. Then there exists a constant $C$ such that for any $\mathbb{P} \in \mathcal{P}_{\alpha,\beta} \cup \mathcal{P}_\beta$ and $\delta \in (0, \frac{1}{2})$, with probability at least $1-\delta$, it holds that*

$$R_{\text{SPO}}(\hat{g}^n_{\text{SPO+}}; \mathbb{P}) - R^*_{\text{SPO}}(\mathbb{P}) \leq \frac{C\sqrt{\log(1/\delta)}}{n^{1/2}}.$$

Notice that Corollary 3.4.1 improves the rate of convergence to $\mathcal{O}(1/\sqrt{n})$ as compared to the $\mathcal{O}(1/n^{1/4})$ rate of Corollary 3.3.1. This matches the rate for the squared $\ell_2$ surrogate developed by [41] (though their result is in the polyhedral case).

*Proof of Corollaries 3.3.1 and 3.4.1.* Let $b = \sup_{\hat{c}\in\mathcal{H}(\mathcal{X}),c\in\mathcal{C}} \ell(\hat{c}, c) \leq 2D_S \sup_{g\in\mathcal{H},x\in\mathcal{X}} \|g(x)\|_2$. For any $\delta > 0$, with probability at least $1-\delta$, for all $g \in \mathcal{H}$, it holds that

$$\left|R_\ell(g; \mathbb{P}) - \hat{R}^n_\ell(g)\right| \leq 4\sqrt{2}D_S\mathfrak{R}^n(\mathcal{H}) + b\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Since $\mathfrak{R}^n(\mathcal{H}) \leq \frac{C'}{\sqrt{n}}$ and $\log(1/\delta) \geq \log(2)$, we know that there exists some universal constant $C_1$ such that

$$4\sqrt{2}D_S\mathfrak{R}^n(\mathcal{H}) + b\sqrt{\frac{2\log(1/\delta)}{n}} \leq C_1\sqrt{\frac{\log(1/\delta)}{n}},$$

for all $\delta \in (0, \frac{1}{2})$ and $n \geq 1$. Since $\hat{g}^n_{\text{SPO+}}$ minimizes the empirical SPO+ risk $\hat{R}^n_{\text{SPO+}}(\cdot)$, we have $\hat{R}^n_{\text{SPO+}}(\hat{g}^n_{\text{SPO+}}) \leq \hat{R}^n_{\text{SPO+}}(g^*_{\text{SPO+}})$. and therefore, with probability at least $1-\delta$, it holds that

$$R_{\text{SPO+}}(\hat{g}^n_{\text{SPO+}}) - R^*_{\text{SPO+}} \leq 2C_1\sqrt{\frac{\log(1/\delta)}{n}}.$$

Recall Theorem 3.3.1, the biconjugate of $\min\{\frac{\epsilon^2}{D_S M}, \epsilon\}$ is $\frac{\epsilon^2}{D_S M}$ for $\epsilon \in [0, \frac{D_S M}{2}]$ and $\epsilon - \frac{D_S M}{4}$ for $\epsilon \in [\frac{D_S M}{2}, \infty]$. Then if the assumption in Corollary 3.3.1 holds, with probability at least $1-\delta$, it holds that

$$R_{\text{SPO}}(\hat{g}^n_{\text{SPO+}}; \mathbb{P}) - R^*_{\text{SPO}}(\mathbb{P}) \leq \frac{C_2\sqrt{\log(1/\delta)}}{n^{1/4}},$$

for some universal constant $C_2$. Also, since the calibration function in Theorem 3.4.1 is linear and thus convex, then if the assumption in Corollary 3.3.1 holds, with probability at least $1-\delta$, it holds that

$$R_{\text{SPO}}(\hat{g}^n_{\text{SPO+}}; \mathbb{P}) - R^*_{\text{SPO}}(\mathbb{P}) \leq \frac{C_3\sqrt{\log(1/\delta)}}{n^{1/2}},$$

for some universal constant $C_3$. □

In the remaining part of this section, we provide the proof of Theorem 3.4.1, Lemma 3.4.1, and Lemma 3.4.2. Also, for any vector $c \in \mathbb{R}^d$, we will use $\tilde{c}$ to represent the projection $\text{proj}_{H^\perp}(c)$ for simplicity. Likewise, when $c = \nabla f(w)$ we shorten this notation even further to $\tilde{\nabla} f(w)$.

First we provide some useful properties in the following lemma.

**Lemma 3.4.3.** *If $f(\cdot)$ is $\mu$-strongly convex on $S$, the for all $w \in S$, it holds that*

$$\|\tilde{\nabla} f(w)\|_*^2 \geq \sqrt{2\mu(f(w) - f_{\min})}.$$

*Proof.* First, for all $c \in \mathbb{R}^d$ and $w, w' \in S$, it holds that

$$c^T(w - w') - \tilde{c}^T(w - w') = (c - \tilde{c})^T(w - w') = \sum_{j=1}^{m_2} \alpha_j h_j(w - w') = 0.$$

Since $f(\cdot)$ is $\mu$-strongly convex, it holds that $f(w') \geq f(x) + \nabla f(w)^T(w' - w) + \frac{\mu}{2}\|w' - w\|^2$ for all $w' \in S$. Therefore, it holds that

$$\begin{aligned}
\inf_{w' \in S} f(w') &\geq \inf_{w' \in S} \left\{ f(w) + \nabla f(w)^T(w' - w) + \frac{\mu}{2}\|w' - w\|^2 \right\} \\
&= \inf_{w' \in S} \left\{ f(w) + \tilde{\nabla} f(w)^T(w' - w) + \frac{\mu}{2}\|w' - w\|^2 \right\} \\
&\geq \inf_{w' \in \mathbb{R}^d} \left\{ f(w) + \tilde{\nabla} f(w)^T(w' - w) + \frac{\mu}{2}\|w' - w\|^2 \right\} \\
&= f(w) - \frac{1}{2\mu}\|\tilde{\nabla} f(w)\|_*^2.
\end{aligned}$$

□

**Lemma 3.4.4.** *If $f(\cdot)$ is $L$-smooth on $S$, then for all $w \in S$, it holds that*

$$\|\tilde{\nabla} f(w)\|_*^2 \leq \sqrt{2L(f(w) - f_{\min})}.$$

*Proof.* If $\tilde{\nabla} f(w) = 0$, then the statement holds. Otherwise, there exists $u \in \mathbb{R}^d$ such that $\|u\| = 1$ and $\tilde{\nabla} f(w)^T u = \|\tilde{\nabla} f(w)\|_*$. Let $v = \|\tilde{\nabla} f(w)\|_* u$, we have $\|v\| = \|\tilde{\nabla} f(w)\|_*$ and $\tilde{\nabla} f(w)^T v = \|\tilde{\nabla} f(w)\|_*^2$. Let

$$\alpha = \sup \alpha', \quad \text{s.t. } f(w - \alpha'\tilde{v}) \leq r.$$

Since $g_i(\cdot)$ is continuous and $g_i(w) < 0$ for all $i \in [m_1]$, we have $\alpha > 0$ and since $f(\cdot)$ is continuous, we have $f(w - \alpha\tilde{v}) = r$. Since $f(\cdot)$ is $L$-smooth on $S$, it holds that

$$\begin{aligned}
f(w - \alpha\tilde{v}) &\leq f(w) - \alpha \nabla f(w)^T \tilde{v} + \frac{\alpha^2 L}{2}\|\tilde{v}\|^2 = f(w) - \alpha \tilde{\nabla} f(w)^T \tilde{v} + \frac{\alpha^2 L}{2}\|\tilde{v}\|^2 \\
&= f(w) - \alpha \tilde{\nabla} f(w)^T v + \frac{\alpha^2 L}{2}\|\tilde{v}\|^2 \leq f(w) - \alpha \tilde{\nabla} f(w)^T v + \frac{\alpha^2 L}{2}\|v\|^2 \\
&= f(w) - \frac{2\alpha - \alpha^2 L}{2}\|\tilde{\nabla} f(w)\|_*^2.
\end{aligned}$$

Therefore, we have $2\alpha - \alpha^2 L \leq 0$. Moreover, since $\alpha > 0$, then it holds that $\alpha \geq \frac{2}{L}$. Now we know that $w - \frac{\tilde{v}}{L} \in S$, and

$$f_{\min} \leq f\left(w - \frac{\tilde{v}}{L}\right) \leq f(w) - \frac{1}{2L}\|\tilde{\nabla} f(w)\|_*^2.$$

Therefore, it holds that $\|\tilde{\nabla} f(w)\|_* \leq \sqrt{2L(f(w) - f_{\min})}$. □

Now we provide the proofs of Lemma 3.4.1 and Lemma 3.4.2.

*Proof of Lemma 3.4.1.* Let $w_1 = w^*(c_1)$ and $w_2 = w^*(c_2)$. Since $f(\cdot)$ is $\mu$-strongly convex on $S$, it holds that

$$f(w) - f(w_1) - \nabla f(w_1)^T(w - w_1) \geq \frac{\mu}{2}\|w - w_1\|^2.$$

Since the Slater condition holds, the KKT necessary condition indicates that there exists scalar $u \geq 0$ such that $\tilde{c}_1 + u\tilde{\nabla} f(w_1) = 0$ and $u(f(w_1) - r) = 0$. When $\tilde{c}_1 \neq 0$, we additionally have $f(w_1) = r$. Therefore, it holds that

$$\begin{aligned} c_1^T(w - w_1) = \tilde{c}_1^T(w - w_1) &= u \cdot \left(-\tilde{\nabla} f(w_1)^T(w - w_1)\right) = u \cdot \left(-\nabla f(w_1)^T(w - w_1)\right) \\ &\geq u \cdot \left(f(w_1) - f(w) + \frac{\mu}{2}\|w - w_1\|^2\right) \geq \frac{u\mu}{2}\|w - w_1\|^2, \end{aligned}$$

where the last inequality holds since $f(w_1) = r \geq f(w)$. Therefore, it holds that

$$c_1^T(w - w_1) \geq \frac{\mu\|\tilde{c}_1\|_*\|w - w_1\|^2}{2\|\tilde{\nabla} f(w_1)\|_*}. \tag{3.13}$$

Since $f(\cdot)$ is $L$-smooth on $S$, it holds that $\|\tilde{\nabla} f(w_1)\|_* \leq \sqrt{2L(r - f_{\min})}$, and hence we have

$$\frac{\|\tilde{c}_1\|_*}{\|\tilde{\nabla} f(w_1)\|_*} \geq \frac{\|\tilde{c}_1\|_*}{\sqrt{2L(r - f_{\min})}}.$$

By applying the above inequality to (3.13), we can conclude that

$$c^T(w - w_1) \geq \frac{\mu}{2\sqrt{2L(r - f_{\min})}}\|\tilde{c}_1\|_*\|w - w_1\|^2.$$

On the other hand, it holds that

$$\begin{aligned} c_1^T(w_2 - w_1) = \tilde{c}_1^T(w_2 - w_1) &= u \cdot \left(-\tilde{\nabla} f(w_1)^T(w_2 - w_1)\right) = u \cdot \left(-\nabla f(w_1)^T(w_2 - w_1)\right) \\ &\leq u \cdot \left(f(w_1) - f(w_2) + \frac{L}{2}\|w_2 - w_1\|^2\right) = \frac{uL}{2}\|w - w_1\|^2, \end{aligned}$$

where the last inequality holds since $f(w_1) = r = f(w_2)$. Therefore, it holds that

$$c_1^T(w_2 - w_1) \leq \frac{L\|\tilde{c_1}\|_*\|w_2 - w_1\|^2}{2\|\tilde{\nabla} f(w_1)\|_*}. \tag{3.14}$$

Since $f(\cdot)$ is $\mu$-strongly convex on $S$, it holds that $\|\tilde{\nabla} f(w_1)\|_* \geq \sqrt{2\mu(r - f_{\min})}$, and hence we have

$$\frac{\|\tilde{c}_1\|_*}{\|\tilde{\nabla} f(w_1)\|_*} \leq \frac{\|\tilde{c}_1\|_*}{\sqrt{2\mu(r - f_{\min})}}.$$

By applying the above inequality to (3.14), we can conclude that

$$c^T(w - w_1) \leq \frac{L}{2\sqrt{2\mu(r - f_{\min})}}\|\tilde{c}_1\|_*\|w_2 - w_1\|^2.$$

□

*Proof of Lemma 3.4.2.* Without loss of generality we assume $\|\tilde{c}_1\|_* = \|\tilde{c}_2\|_* = 1$. Let $w_1 = w^*(c_1)$ and $w_2 = w^*(c_2)$. By KKT condition there exists $u_1, u_2 > 0$ such that $\nabla f(w_i) = -u_i c_i$ and $f(w_i) = r$ for $i = 1, 2$. Also, since $f(\cdot)$ is $\mu$-strongly convex, it holds that

$$\|\tilde{\nabla} f(w_i)\|_* \geq \sqrt{2\mu(f(x_i) - f_{\min})} = \sqrt{2\mu(r - f_{\min})},$$

for $i = 1, 2$. Then, it holds that

$$\|\tilde{\nabla} f(w_1) - \tilde{\nabla} f(w_2)\|_* \geq \min_{u_1', u_2' \geq \sqrt{2\mu(r - f_{\min})}} \|u_1'\tilde{c}_1 - u_2'\tilde{c}_2\|_* = \sqrt{2\mu(r - f_{\min})} \cdot \|\tilde{c}_1 - \tilde{c}_2\|_*.$$

Moreover, since $f(\cdot)$ is $L$-smooth, it holds that

$$\|w_1 - w_2\| \geq \frac{1}{L} \cdot \|\nabla f(w_1) - \nabla f(w_2)\|_* \geq \frac{1}{L} \cdot \|\tilde{\nabla} f(w_1) - \tilde{\nabla} f(w_2)\|_* \geq \frac{\sqrt{2\mu r}}{L} \cdot \|\tilde{c}_1 - \tilde{c}_2\|_*.$$

□

In the rest part of this section, without loss of generality we assume $f_{\min} = 0$. Also, since $w*(c) = w^*(\tilde{c})$ and $c^T(w^*(c') - w^*(c)) = \tilde{c}^T(w^*(c') - w^*(c))$ for all $c, c' \in \mathbb{R}^d$, we will ignore the $\tilde{}$ notation and assume all $c, c' \in H^\perp$. In Theorem 3.4.2 we provide a lower bound of an SPO-like loss.

**Theorem 3.4.2.** *When $c \neq 0$ and $c + 2\Delta \neq 0$, it holds that*

$$(c + 2\Delta)^T(w^*(c) - w^*(c + 2\Delta)) \geq \frac{\mu^2 r^{1/2}}{2^{1/2} L^{5/2}} \cdot \|c + 2\Delta\|_* \cdot \left\| \frac{c}{\|c\|_*} - \frac{c + 2\Delta}{\|c + 2\Delta\|_*} \right\|_*^2.$$

*When the norm we consider is A-norm defined by $\|x\|_A = \sqrt{x^T A x}$ for some positive definite matrix A, additionally we have*

$$(c+2\Delta)^T(w^*(c)-w^*(c+2\Delta)) \geq \frac{\mu^2 r^{1/2}}{2^{1/2}L^{5/2}}\left(\|c+2\Delta\|_{A^{-1}} - \frac{c^T A^{-1}(c+2\Delta)}{\|c\|_{A^{-1}}}\right).$$

*Moreover, if $\mathbb{P}(c=0) = \mathbb{P}(c=-2\Delta) = 0$, it holds that*

$$\ell_{\text{SPO+}}(\Delta) \geq \frac{\mu^2 r^{1/2}}{2^{1/2}L^{5/2}} \cdot \mathbb{E}_c\left[\|c+2\Delta\|_{A^{-1}} - \frac{c^T A^{-1}(c+2\Delta)}{\|c\|_{A^{-1}}}\right].$$

*Proof of Theorem 3.4.2.* Apply $c_1 = c$ and $c_2 = c+2\Delta$ to Lemma 3.4.2, we have

$$\|w^*(c)-w^*(c+2\Delta)\| \geq \sqrt{2\mu r}\cdot\left\|\frac{c}{\|c\|_*} - \frac{c+2\Delta}{\|c+2\Delta\|_*}\right\|_*.$$

By applying the above inequality to (3.11) we have

$$\begin{aligned}(c+2\Delta)^T(w^*(c)-w^*(c+2\Delta)) &\geq \frac{\mu}{2\sqrt{2Lr}}\cdot\|c+2\Delta\|_*\cdot\|w^*(c)-w^*(c+2\Delta)\|^2\\ &\geq \frac{\mu}{2\sqrt{2Lr}}\cdot\|c+2\Delta\|_*\cdot\left(\frac{\sqrt{2\mu r}}{L}\cdot\left\|\frac{c}{\|c\|_*} - \frac{c+2\Delta}{\|c+2\Delta\|_*}\right\|_*\right)^2\\ &= \frac{\mu^2 r^{1/2}}{2^{1/2}L^{5/2}}\cdot\|c+2\Delta\|_*\cdot\left\|\frac{c}{\|c\|_*} - \frac{c+2\Delta}{\|c+2\Delta\|_*}\right\|_*^2.\end{aligned}$$

When the norm we consider is $A$-norm, then it holds that

$$\begin{aligned}(c+2\Delta)^T(w^*(c)-w^*(c+2\Delta)) &\geq \frac{\mu^2 r^{1/2}}{2^{1/2}L^{5/2}}\cdot\|c+2\Delta\|_2\cdot\left\|\frac{c}{\|c\|_{A^{-1}}} - \frac{c+2\Delta}{\|c+2\Delta\|_{A^{-1}}}\right\|_{A^{-1}}^2\\ &= \frac{\mu^2 r^{1/2}}{2^{1/2}L^{5/2}}\left(\|c+2\Delta\|_{A^{-1}} - \frac{c^T A^{-1}(c+2\Delta)}{\|c\|_{A^{-1}}}\right).\end{aligned}$$

Moreover, if $\mathbb{P}(c=0) = \mathbb{P}(c=-2\Delta) = 0$, by taking the expectation of $c$ we get

$$\ell_{\text{SPO+}}(\Delta) \geq \frac{\mu^2 r^{1/2}}{2^{1/2}L^{5/2}} \cdot \mathbb{E}_c\left[\|c+2\Delta\|_{A^{-1}} - \frac{c^T A^{-1}(c+2\Delta)}{\|c\|_{A^{-1}}}\right].$$

□

The following lemma provides a necessary condition on $\Delta$ such that the excess SPO loss of $\hat{c} = \bar{c}+\Delta$ is at least $\epsilon$.

**Lemma 3.4.5.** *Suppose the excess SPO loss of $\hat{c} = \bar{c}+\Delta$ is at least $\epsilon$, that is, $\bar{c}^T(w^*(\bar{c}+\Delta) - w^*(\bar{c})) \geq \epsilon$. Then it holds that*

$$\left\|\frac{\bar{c}}{\|\bar{c}\|_*} - \frac{\bar{c}+\Delta}{\|\bar{c}+\Delta\|_*}\right\|_*^2 \geq \frac{2^{1/2}\mu^{5/2}\epsilon}{L^2 r^{1/2}\|\bar{c}\|_*}.$$

*When the norm we consider is A-norm defined by $\|x\|_A = \sqrt{x^T A x}$ for some positive definite matrix A, additionally we have*

$$1 - \frac{\bar{c}^T A^{-1}(\bar{c}+\Delta)}{\|\bar{c}\|_{A^{-1}} \cdot \|\bar{c}+\Delta\|_{A^{-1}}} \geq \frac{\mu^{5/2}}{2^{1/2} L^2 r^{1/2} \|\bar{c}\|_{A^{-1}}} \cdot \epsilon.$$

*Proof of Lemma 3.4.5.* In Lemma 3.4.1 we show that

$$c_1^T(w^*(c_2) - w^*(c_1)) \leq \frac{L}{2\sqrt{2\mu r}} \|c_1\|_* \|w^*(c_1) - w^*(c_2)\|^2.$$

Let $c_1 = \bar{c}$ and $c_2 = \hat{c}$, it holds that

$$\|w^*(\bar{c}) - w^*(\bar{c}+\Delta)\|^2 \geq \frac{2\sqrt{2\mu r}}{L\|\bar{c}\|_*} \cdot \bar{c}^T(w^*(\bar{c}+\Delta) - w^*(\bar{c})) \geq \frac{2\sqrt{2\mu r}\epsilon}{L\|\bar{c}\|_*}.$$

Theorem 3 in [28] shows that for $c_1, c_2 \in \mathbb{R}^d$, it holds that

$$\|c_1 - c_2\|_* \geq \frac{\mu}{\sqrt{2Lr}} \cdot \min\{\|c_1\|_*, \|c_2\|_*\} \cdot \|w^*(c_1) - w^*(c_2)\|.$$

By applying $c_1 = \frac{\bar{c}}{\|\bar{c}\|_*}$ and $c_2 = \frac{\bar{c}+\Delta}{\|\bar{c}+\Delta\|_*}$, we have

$$\begin{aligned}\left\| \frac{\bar{c}}{\|\bar{c}\|_*} - \frac{\bar{c}+\Delta}{\|\bar{c}+\Delta\|_*} \right\|^2 &\geq \frac{\mu^2}{2Lr} \cdot \left\| w^*\left(\frac{\bar{c}}{\|\bar{c}\|_*}\right) - w^*\left(\frac{\bar{c}+\Delta}{\|\bar{c}+\Delta\|_*}\right) \right\|^2 \\ &= \frac{\mu^2}{2Lr} \cdot \|w^*(\bar{c}) - w^*(\bar{c}+\Delta)\|_*^2 \geq \frac{2^{1/2}\mu^{5/2}\epsilon}{L^2 r^{1/2}\|\bar{c}\|_*}.\end{aligned}$$

When the norm we consider is 2-norm, it holds that

$$1 - \frac{\bar{c}^T A^{-1}(\bar{c}+\Delta)}{\|\bar{c}\|_{A^{-1}} \cdot \|\bar{c}+\Delta\|_{A^{-1}}} = \frac{1}{2} \left\| \frac{\bar{c}}{\|\bar{c}\|_{A^{-1}}} - \frac{\bar{c}+\Delta}{\|\bar{c}+\Delta\|_{A^{-1}}} \right\|_{A^{-1}}^2 \geq \frac{\mu^{5/2}\epsilon}{2^{1/2} L^2 r^{1/2} \|\bar{c}\|_{A^{-1}}}.$$

□

From Theorem 3.4.2 and Lemma 3.4.5, we know that $\ell_{\text{SPO+}}(c, \Delta)$ have a lower bound $C_1(\mu, L, r) \cdot \underline{\ell}_{\text{SPO+}}(c, \Delta)$, where $C_1(\mu, L, r) = \frac{\mu^2 r^{1/2}}{2^{1/2} L^{5/2}}$ and

$$\underline{\ell}_{\text{SPO+}}(c, \Delta) = \|c + 2\Delta\|_{A^{-1}} - \frac{c^T A^{-1}(c+2\Delta)}{\|c\|_{A^{-1}}}.$$

Moreover, the excess SPO risk of $\hat{c} = \bar{c} + \Delta$ is at least $\epsilon$ implies that $\overline{R}_{\text{SPO}}(\Delta) \geq C_2(\mu, L, r) \cdot \epsilon$ where $C_2(\mu, L, r) = \frac{\mu^{5/2}}{2^{1/2} L^2 r^{1/2}}$ and

$$\overline{R}_{\text{SPO}}(\Delta) = \|\bar{c}\|_{A^{-1}} - \frac{\bar{c}^T A^{-1}(\bar{c}+\Delta)}{\|\bar{c}+\Delta\|_{A^{-1}}}.$$

Let $\underline{R}_{\text{SPO+}}(\Delta) = \mathbb{E}_c[\underline{\ell}_{\text{SPO+}}(c, \Delta)]$. We know that the calibration function $\delta(\epsilon)$ has a lower bound $\delta'(\epsilon)$ which defined as

$$\begin{aligned} \delta'(\epsilon) := \min_{\Delta} \quad & C_1(\mu, L, r) \cdot \underline{R}_{\text{SPO+}}(\Delta) \\ \text{s.t.} \quad & \overline{R}_{\text{SPO}}(\Delta) \geq C_2(\mu, L, r) \cdot \epsilon. \end{aligned} \tag{3.15}$$

Here we first provide two properties of random variable $c$ when $\mathbb{P} \in \mathcal{P}_{\text{rot symm}}$.

**Proposition 3.4.1.** *Suppose $\mathbb{P} \in \mathcal{P}_{\text{rot symm}}$. If $\|\bar{c} + \zeta\|_{A^{-1}} = \|\bar{c}\|_{A^{-1}}$ for some $\zeta \in \mathbb{R}^p$, it holds that*

$$\mathbb{E}_c \left[ \|c + \zeta\|_{A^{-1}} \right] = \mathbb{E}_c \left[ \|c\|_{A^{-1}} \right].$$

**Proposition 3.4.2.** *Suppose $\mathbb{P} \in \mathcal{P}_{\text{rot symm}}$. When $d \geq 2$, for any constant $t \geq 0$, it holds that*

$$\mathbb{E}_c \left[ \frac{\bar{c}^T A^{-1} c}{\|c\|_{A^{-1}}} \middle| \|c - \bar{c}\|_{A^{-1}} = t \right] \geq \frac{\|c\|_{A^{-1}}^2 \min\{\|c\|_{A^{-1}}, t\}}{t^2 + \|\bar{c}\|_{A^{-1}} t}.$$

*Proof of Proposition 3.4.2.* For simplicity we just assume $\|c - \bar{c}\|_{A^{-1}} = t$ from now on and ignore the conditional probability. Let $\omega = c - \bar{c}$. Since $p(c) = p(2\bar{c} - c)$, we have

$$\begin{aligned} \mathbb{E}_c \left[ \frac{\bar{c}^T A^{-1} c}{\|c\|_{A^{-1}}} \right] &= \frac{1}{2} \cdot \mathbb{E}_c \left[ \frac{\bar{c}^T A^{-1} c}{\|c\|_{A^{-1}}} + \frac{\bar{c}^T A^{-1}(2\bar{c} - c)}{\|2\bar{c} - c\|_{A^{-1}}} \right] \\ &= \frac{1}{2} \cdot \mathbb{E}_\omega \left[ \frac{\bar{c}^T A^{-1}(\bar{c} + \omega)}{\|\bar{c} + \omega\|_{A^{-1}}} + \frac{\bar{c}^T A^{-1}(\bar{c} - \omega)}{\|\bar{c} - \omega\|_{A^{-1}}} \right]. \end{aligned}$$

By the fact that $\bar{c}^T A^{-1} \bar{c}(\|\bar{c} - w\|_{A^{-1}} + \|\bar{c} + w\|_{A^{-1}}) \geq 2\|\bar{c}\|_{A^{-1}}^2 \|w\|_{A^{-1}} \geq \bar{c}^T A^{-1} w(\|\bar{c} - w\|_{A^{-1}} - \|\bar{c} + w\|_{A^{-1}})$, it holds that $\bar{c}^T A^{-1}(\bar{c} + w)\|\bar{c} - w\|_{A^{-1}} + \bar{c}^T A^{-1}(\bar{c} - w)\|\bar{c} - w\|_{A^{-1}} \geq 0$ and hence

$$\frac{\bar{c}^T A^{-1}(\bar{c} + \omega)}{\|\bar{c} + \omega\|_{A^{-1}}} + \frac{\bar{c}^T A^{-1}(\bar{c} - \omega)}{\|\bar{c} - \omega\|_{A^{-1}}} \geq 0.$$

Therefore, we further get

$$\mathbb{E}_c \left[ \frac{\bar{c}^T A^{-1} c}{\|c\|_{A^{-1}}} \right] \geq \frac{1}{2} \cdot \mathbb{E}_\omega \left[ \frac{\bar{c}^T A^{-1}(\bar{c} + \omega)}{\|\bar{c} + \omega\|_{A^{-1}}} + \frac{\bar{c}^T A^{-1}(\bar{c} - \omega)}{\|\bar{c} - \omega\|_{A^{-1}}} \middle| \bar{c}^T \omega \in C \right] \cdot \mathbb{P}(\bar{c}^T \omega \in C),$$

where $C = [-\|\bar{c}\|_{A^{-1}}^2, \|\bar{c}\|_{A^{-1}}^2]$. For any $\omega$ such that $\bar{c}^T \omega \in C$, we have

$$\begin{aligned} \frac{\bar{c}^T A^{-1}(\bar{c} + \omega)}{\|\bar{c} + \omega\|_{A^{-1}}} + \frac{\bar{c}^T A^{-1}(\bar{c} - \omega)}{\|\bar{c} - \omega\|_{A^{-1}}} &\geq \frac{\bar{c}^T A^{-1}(\bar{c} + \omega)}{\|\bar{c}\|_{A^{-1}} + \|\omega\|_{A^{-1}}} + \frac{\bar{c}^T A^{-1}(\bar{c} - \omega)}{\|\bar{c}\|_{A^{-1}} + \|\omega\|_{A^{-1}}} \\ &= \frac{2\bar{c}^T A^{-1} \bar{c}}{\|\bar{c}\|_{A^{-1}} + \|\omega\|_{A^{-1}}}. \end{aligned}$$

Also, when $d \geq 2$, we have $\mathbb{P}(\bar{c}^T \omega \in C) \geq \frac{\min\{\|\bar{c}\|, t\}}{t}$. Then we can conclude that

$$\mathbb{E}_c \left[ \frac{\bar{c}^T A^{-1} c}{\|c\|_{A^{-1}}} \right] \geq \frac{\|\bar{c}\|_{A^{-1}}^2 \min\{\|\bar{c}\|_{A^{-1}}, t\}}{t^2 + \|\bar{c}\|_{A^{-1}} t}.$$

□

By first-order necessary condition we know that $\Delta$ is an optimal solution to (3.15) only if

$$\nabla \underline{R}_{\mathrm{SPO+}}(\Delta) - \alpha \nabla \overline{R}_{\mathrm{SPO}}(\Delta) = 0 \tag{3.16}$$

for some $\alpha \geq 0$. Also, for any fixed $\Delta$, it holds that

$$\nabla \underline{R}_{\mathrm{SPO+}}(\Delta) = \mathbb{E}_c \left[ \frac{A^{-1}(c+2\Delta)}{\|c+2\Delta\|_{A^{-1}}} - \frac{A^{-1}c}{\|c\|_{A^{-1}}} \right],$$

and

$$\nabla \overline{R}_{\mathrm{SPO}}(\Delta) = \frac{\bar{c}^T A^{-1}(\bar{c}+\Delta) \cdot A^{-1}\Delta - \Delta^T A^{-1}(\bar{c}+\Delta) \cdot A^{-1}\bar{c}}{\|\bar{c}+\Delta\|^3_{A^{-1}}}.$$

The following lemma simplifies $\nabla \ell_{\mathrm{SPO+}}(\Delta)$.

**Lemma 3.4.6.** *Suppose $\mathbb{P} \in \mathcal{P}_{\text{rot symm}}$. Then there exists a unique function $\zeta(\cdot) : [0,\infty] \to [0,\infty]$ such that for all $\Delta \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}_c \left[ \frac{c+\Delta}{\|c+\Delta\|_{A^{-1}}} \right] = \zeta(\|\bar{c}+\Delta\|_{A^{-1}})(\bar{c}+\Delta).$$

*Also, $\alpha \cdot \zeta(\alpha)$ is a non-decreasing function.*

*Proof.* Let $h(\Delta)$ denote $\mathbb{E}_c[\frac{c+\Delta}{\|c+\Delta\|}]$. First we show that $h(\Delta)$ has the same direction as $\bar{c}+\Delta$. Let $\phi_\Delta(\cdot)$ denote the affine transform $\phi_\Delta(\cdot) : \xi \to \frac{2(\bar{c}+\Delta)^T A^{-1}\xi}{\|\bar{c}+\Delta\|^2_{A^{-1}}}(\bar{c}+\Delta) - \xi$. We have $\phi_\Delta(\phi_\Delta(\xi)) = \xi$ and $\|\xi\|_{A^{-1}} = \|\phi_\Delta(\xi)\|_{A^{-1}}$ for all $\xi \in \mathbb{R}^d$. It leads to $p(\xi) = p(\phi_\Delta(\xi))$ and hence

$$\begin{aligned} h(\Delta) &= \frac{1}{2}\mathbb{E}_c \left[ \frac{c+\Delta}{\|c+\Delta\|_{A^{-1}}} + \frac{\phi_\Delta(c+\Delta)}{\|\phi_\Delta(c+\Delta)\|_{A^{-1}}} \right] = \frac{1}{2}\mathbb{E}_c \left[ \frac{(c+\Delta)+\phi_\Delta(c+\Delta)}{\|c+\Delta\|_{A^{-1}}} \right] \\ &= \mathbb{E}_c \left[ \frac{(\bar{c}+\Delta)^T A^{-1}(c+\Delta)}{\|c+\Delta\|_{A^{-1}} \cdot \|\bar{c}+\Delta\|^2_{A^{-1}}} \right] (\bar{c}+\Delta). \end{aligned}$$

Now we let

$$\hat{\zeta}(\bar{c}+\Delta) = \mathbb{E}_c \left[ \frac{(\bar{c}+\Delta)^T A^{-1}(c+\Delta)}{\|c+\Delta\|_{A^{-1}} \cdot \|\bar{c}+\Delta\|^2_{A^{-1}}} \right],$$

and we want to show that $\hat{\zeta}(\bar{c}+\Delta) = \hat{\zeta}(\bar{c}+\Delta')$ if $\|\bar{c}+\Delta\|_{A^{-1}} = \|\bar{c}+\Delta'\|_{A^{-1}}$. Since $\|\bar{c}+\Delta\|_{A^{-1}} = \|\bar{c}+\Delta'\|_{A^{-1}}$, there exists a matrix $R \in \mathbb{R}^{d\times d}$ such that $A^{-1/2}(\bar{c}+\Delta') = RA^{-1/2}(\bar{c}+\Delta)$ and $RR^T = R^T R = I$. Let $c'$ be a random variable depending on $c$ where $c' = A^{1/2}RA^{-1/2}(c-\bar{c}) + \bar{c}$. It holds that $A^{-1/2}(c'-\bar{c}) = RA^{-1/2}(c-\bar{c})$, which implies that $\|c'-\bar{c}\|_{A^{-1}} = \|c-\bar{c}\|_{A^{-1}}$ and therefore $p(c-\bar{c}) = p(c'-\bar{c})$. Also, we have $A^{-1/2}(c'+\Delta') = RA^{-1/2}(c+\Delta)$, which implies that $\|c'+\Delta'\|_{A^{-1}} = \|c+\Delta\|_{A^{-1}}$ and therefore

$$\frac{(\bar{c}+\Delta')^T A^{-1}(c'+\Delta')}{\|c'+\Delta'\|_{A^{-1}}} = \frac{(\bar{c}+\Delta)^T A^{-1/2} R^T R A^{-1/2}(c+\Delta)}{\|c+\Delta\|_{A^{-1}}} = \frac{(\bar{c}+\Delta)^T A^{-1}(c+\Delta)}{\|c+\Delta\|_{A^{-1}}}.$$

Moreover, since $\det(A^{1/2}RA^{-1/2}) = 1$, it holds that

$$\mathbb{E}_c\left[\frac{(\bar{c}+\Delta')^T A^{-1}(c+\Delta')}{\|c+\Delta'\|_{A^{-1}}}\right] = \mathbb{E}_c\left[\frac{(\bar{c}+\Delta')^T A^{-1}(c'+\Delta')}{\|c'+\Delta'\|_{A^{-1}}}\right] = \mathbb{E}_c\left[\frac{(\bar{c}+\Delta')^T A^{-1}(c+\Delta')}{\|c+\Delta'\|_{A^{-1}}}\right].$$

Therefore,

$$\begin{aligned}\hat{\zeta}(\bar{c}+\Delta) &= \frac{1}{\|\bar{c}+\Delta\|_{A^{-1}}^2}\cdot\mathbb{E}_c\left[\frac{(\bar{c}+\Delta')^T A^{-1}(c+\Delta')}{\|c+\Delta'\|_{A^{-1}}}\right] \\ &= \frac{1}{\|\bar{c}+\Delta'\|_{A^{-1}}^2}\cdot\mathbb{E}_c\left[\frac{(\bar{c}+\Delta')^T A^{-1}(c+\Delta')}{\|c+\Delta'\|_{A^{-1}}}\right] = \hat{\zeta}(\bar{c}+\Delta').\end{aligned}$$

Therefore, we know that $\zeta(\cdot):\mathbb{R}\to\mathbb{R}$ is a well-defined function based on the above property of $\hat{\zeta}(\cdot)$. Now we are going to prove that $\alpha\cdot\zeta(\alpha)$ is a non-decreasing function. Pick arbitrary $\alpha_1' > \alpha_2' > 0$, we have $\zeta(\alpha_1') = \hat{\zeta}(\alpha_1\cdot\bar{c})$ and $\zeta(\alpha_2') = \hat{\zeta}(\alpha_2\cdot\bar{c})$, where $\alpha_i = \alpha_i'/\|\bar{c}\|_{A^{-1}}$ for $i = 1, 2$. Therefore,

$$\begin{aligned}&\alpha_1'\cdot\zeta(\alpha_1') \geq \alpha_2'\cdot\zeta(\alpha_2') \Leftrightarrow \alpha_1\cdot\hat{\zeta}(\alpha_1\cdot\bar{c}) \geq \alpha_2\cdot\hat{\zeta}(\alpha_2\cdot\bar{c}) \\ \Leftrightarrow&\alpha_1\mathbb{E}_c\left[\frac{(\alpha_1\cdot\bar{c})^T A^{-1}((c-\bar{c})+\alpha_1\cdot\bar{c})}{\|(c-\bar{c})+\alpha_1\cdot\bar{c}\|_{A^{-1}}\cdot\|\alpha_1\cdot\bar{c}\|_{A^{-1}}^2}\right] \geq \alpha_2\mathbb{E}_c\left[\frac{(\alpha_2\cdot\bar{c})^T A^{-1}((c-\bar{c})+\alpha_2\cdot\bar{c})}{\|(c-\bar{c})+\alpha_2\cdot\bar{c}\|_{A^{-1}}\cdot\|\alpha_2\cdot\bar{c}\|_{A^{-1}}^2}\right] \\ \Leftrightarrow&\mathbb{E}_c\left[\frac{\bar{c}^T A^{-1}((c-\bar{c})+\alpha_1\cdot\bar{c})}{\|(c-\bar{c})+\alpha_1\cdot\bar{c}\|_{A^{-1}}}\right] \geq \mathbb{E}_c\left[\frac{\bar{c}^T A^{-1}((c-\bar{c})+\alpha_2\cdot\bar{c})}{\|(c-\bar{c})+\alpha_2\cdot\bar{c}\|_{A^{-1}}}\right].\end{aligned}$$

It is sufficient to show that

$$\frac{\bar{c}^T A^{-1}(\zeta+\alpha_1\cdot\bar{c})}{\|\zeta+\alpha_1\cdot\bar{c}\|_{A^{-1}}} \geq \frac{\bar{c}^T A^{-1}(\zeta+\alpha_2\cdot\bar{c})}{\|\zeta+\alpha_2\cdot\bar{c}\|_{A^{-1}}}, \tag{3.17}$$

for all $\zeta\in\mathbb{R}^d$ when $\alpha_1 > \alpha_2 > 0$. We divide the proof into three cases. When $\bar{c}^T A^{-1}(\zeta+\alpha_1\cdot\bar{c}) > \bar{c}^T A^{-1}(\zeta+\alpha_2\cdot\bar{c}) \geq 0$, (3.17) is equivalent to

$$\begin{aligned}&\left(\bar{c}^T A^{-1}(\zeta+\alpha_1\cdot\bar{c})\right)^2\cdot\|\zeta+\alpha_2\cdot\bar{c}\|_{A^{-1}}^2 \geq \left(\bar{c}^T A^{-1}(\zeta+\alpha_2\cdot\bar{c})\right)^2\cdot\|\zeta+\alpha_1\cdot\bar{c}\|_{A^{-1}}^2 \\ \Leftrightarrow&(\alpha_1-\alpha_2)\left(\bar{c}^T A^{-1}(\zeta+\alpha_1\cdot\bar{c})+\bar{c}^T A^{-1}(\zeta+\alpha_2\cdot\bar{c})\right)\left(\bar{c}^T A^{-1}\bar{c}\cdot\zeta^T A^{-1}\zeta-(\bar{c}^T A^{-1}\zeta)^2\right) \geq 0.\end{aligned}$$

When $\bar{c}^T(\zeta+\alpha_1\cdot\bar{c}) \geq 0 \geq \bar{c}^T(\zeta+\alpha_2\cdot\bar{c})$, we know that left hand side of (3.17) is non-negative and right hand side is non-positive. When $0 > \bar{c}^T(\zeta+\alpha_1\cdot\bar{c}) \geq \bar{c}^T(\zeta+\alpha_2\cdot\bar{c})$, (3.17) is equivalent to

$$\begin{aligned}&\left(\bar{c}^T A^{-1}(\zeta+\alpha_1\cdot\bar{c})\right)^2\cdot\|\zeta+\alpha_2\cdot\bar{c}\|_{A^{-1}}^2 \leq \left(\bar{c}^T A^{-1}(\zeta+\alpha_2\cdot\bar{c})\right)^2\cdot\|\zeta+\alpha_1\cdot\bar{c}\|_{A^{-1}}^2 \\ \Leftrightarrow&(\alpha_1-\alpha_2)\left(\bar{c}^T A^{-1}(\zeta+\alpha_1\cdot\bar{c})+\bar{c}^T A^{-1}(\zeta+\alpha_2\cdot\bar{c})\right)\left(\bar{c}^T A^{-1}\bar{c}\cdot\zeta^T A^{-1}\zeta-(\bar{c}^T A^{-1}\zeta)^2\right) \leq 0.\end{aligned}$$

□

Following the results in Lemma 3.4.6 we have

$$\mathbb{E}_c\left[\frac{c}{\|c\|_{A^{-1}}}\right] = \zeta(\|\bar{c}\|_{A^{-1}})\bar{c}, \quad \mathbb{E}_c\left[\frac{c+2\Delta}{\|c+2\Delta\|_{A^{-1}}}\right] = \zeta(\|\bar{c}+2\Delta\|_{A^{-1}})(\bar{c}+2\Delta).$$

Hence, (3.16) is equivalent to

$$\zeta(\|\bar{c}+2\Delta\|_{A^{-1}})(\bar{c}+2\Delta) - \zeta(\|\bar{c}\|_{A^{-1}})\bar{c} = \alpha \cdot \frac{\bar{c}^T A^{-1}(\bar{c}+\Delta)\cdot\Delta - \Delta^T A^{-1}(\bar{c}+\Delta)\cdot\bar{c}}{\|\bar{c}+\Delta\|^3_{A^{-1}}}.$$

Since $\bar{c}$ and $\Delta$ are linearly independent, (3.16) is further equivalent to

$$\frac{2\zeta(\|\bar{c}+2\Delta\|_{A^{-1}})}{\bar{c}^T A^{-1}(\bar{c}+\Delta)} = \frac{\alpha}{\|\bar{c}+2\Delta\|^3_{A^{-1}}} = \frac{\zeta(\|\bar{c}+2\Delta\|_{A^{-1}}) - \zeta(\|\bar{c}\|_{A^{-1}})}{-\Delta^T A^{-1}(\bar{c}+\Delta)},$$

which is also equivalent to

$$(\bar{c}+2\Delta)^T A^{-1}(\bar{c}+\Delta)\cdot\zeta(\|\bar{c}+2\Delta\|_{A^{-1}}) = \bar{c}^T A^{-1}(\bar{c}+\Delta)\cdot\zeta(\|\bar{c}\|_{A^{-1}}). \tag{3.18}$$

**Lemma 3.4.7.** *Suppose $\mathbb{P} \in \mathcal{P}_{\text{rot symm}}$ and $\hat{\Delta}$ is a solution to* (3.16)*, then it holds that*

$$\|\bar{c}+2\hat{\Delta}\|_{A^{-1}} = \|\bar{c}\|_{A^{-1}},$$

*and*

$$(\bar{c}+2\hat{\Delta})^T A^{-1}(\bar{c}+\hat{\Delta}) = \bar{c}^T A^{-1}(\bar{c}+\hat{\Delta}).$$

*Proof.* Suppose $\|\bar{c}+2\hat{\Delta}\|_{A^{-1}} \neq \|\bar{c}\|_{A^{-1}}$. Without loss of generality we assume $\|\bar{c}+2\hat{\Delta}\|_{A^{-1}} > \|\bar{c}\|_{A^{-1}}$. Following results in Lemma 3.4.6 we know that

$$\|\bar{c}+2\Delta\|_{A^{-1}}\cdot\zeta(\|\bar{c}+2\Delta\|_{A^{-1}}) \geq \|\bar{c}\|_{A^{-1}}\cdot\zeta(\|\bar{c}\|_{A^{-1}}).$$

Also, it holds that

$$\hat{\Delta}^T A^{-1}(\bar{c}+\hat{\Delta}) = \frac{1}{4}\left(\|\bar{c}+2\hat{\Delta}\|^2_{A^{-1}} - \|\bar{c}\|^2_{A^{-1}}\right) > 0.$$

Since $(\bar{c}+2\hat{\Delta})^T A^{-1}(\bar{c}+\hat{\Delta}) = (\bar{c}+\hat{\Delta})^T A^{-1}(\bar{c}+\hat{\Delta}) + \hat{\Delta}^T A^{-1}(\bar{c}+\hat{\Delta}) > 0$, it holds that

$$\begin{aligned}
&\frac{(\bar{c}+2\hat{\Delta})^T A^{-1}(\bar{c}+\hat{\Delta})}{\|\bar{c}+2\hat{\Delta}\|_{A^{-1}}} > \frac{\bar{c}^T A^{-1}(\bar{c}+\hat{\Delta})}{\|\bar{c}\|_{A^{-1}}} \\
\Leftrightarrow &(\bar{c}+2\hat{\Delta})^T A^{-1}(\bar{c}+\hat{\Delta})\cdot\|\bar{c}\|_{A^{-1}} > \bar{c}^T A^{-1}(\bar{c}+\hat{\Delta})\cdot\|\bar{c}+2\hat{\Delta}\|_{A^{-1}} \\
\Leftarrow &\left((\bar{c}+2\hat{\Delta})^T A^{-1}(\bar{c}+\hat{\Delta})\right)^2\cdot\|\bar{c}\|^2_{A^{-1}} > \left(\bar{c}^T A^{-1}(\bar{c}+\hat{\Delta})\right)^2\cdot\|\bar{c}+2\hat{\Delta}\|^2_{A^{-1}} \\
\Leftrightarrow &\left(\hat{\Delta}^T A^{-1}(\bar{c}+\hat{\Delta})\right)\cdot\left(\|\bar{c}+\hat{\Delta}\|^2_{A^{-1}}\cdot\|\hat{\Delta}\|^2_{A^{-1}} - (\hat{\Delta}^T A^{-1}(\bar{c}+\hat{\Delta}))^2\right) > 0.
\end{aligned}$$

Therefore, we have

$$(\bar{c}+2\Delta)^T A^{-1}(\bar{c}+\Delta)\cdot\zeta(\|\bar{c}+2\Delta\|_{A^{-1}}) > \bar{c}^T A^{-1}(\bar{c}+\Delta)\cdot\zeta(\|\bar{c}\|_{A^{-1}}),$$

which contradicts with (3.18). Therefore, we have $\|\bar{c}+2\hat{\Delta}\|_{A^{-1}} = \|\bar{c}\|_{A^{-1}}$ and hence $(\bar{c}+2\hat{\Delta})^T A^{-1}(\bar{c}+\hat{\Delta}) = \bar{c}^T A^{-1}(\bar{c}+\hat{\Delta})$. □

Based on the above property, we provide a lower bound of calibration function.

**Theorem 3.4.3.** *Suppose Assumption 3.4.1 holds and $\mathbb{P}\in\mathcal{P}_{\text{rot symm}}$, then the calibration function $\delta(\cdot)$ satisfies*

$$\delta(\epsilon)\geq \mathbb{E}_c\left[\frac{\min\{\|\bar{c}\|_{A^{-1}},\|c-\bar{c}\|_{A^{-1}}\}}{\|c-\bar{c}\|_{A^{-1}}^2+\|\bar{c}\|_{A^{-1}}\|c-\bar{c}\|_{A^{-1}}}\right]\cdot\frac{\mu^{9/2}\|\bar{c}\|_{A^{-1}}}{2L^{9/2}}\cdot\epsilon,$$

*for all $\epsilon>0$.*

*Proof.* First we know that $\delta(\epsilon)\geq\delta'(\epsilon)$. Also, Lemma 3.4.7 shows that for optimal $\Delta$, it holds that $\|\bar{c}\|_{A^{-1}} = \|\bar{c}+2\Delta\|_{A^{-1}}$. By the definition of $\underline{R}_{\text{SPO+}}$, we have

$$\begin{aligned}\underline{R}_{\text{SPO+}}(\Delta) &= \mathbb{E}_c[\underline{\ell}_{\text{SPO+}}(c,\Delta)] = \mathbb{E}_c\left[\|c+2\Delta\|_{A^{-1}} - \frac{c^T A^{-1}(c+2\Delta)}{\|c\|_{A^{-1}}}\right]\\ &= \mathbb{E}_c\left[\|c+2\Delta\|_{A^{-1}}\right] - \mathbb{E}_c\left[\|c\|_{A^{-1}}\right] - \mathbb{E}_c\left[\frac{2c^T A^{-1}\Delta}{\|c\|_{A^{-1}}}\right].\end{aligned}$$

Since $\|\bar{c}+2\Delta\|_{A^{-1}} = \|\bar{c}\|_{A^{-1}}$, Proposition 3.4.1 shows that $\mathbb{E}_c\left[\|c+2\Delta\|_{A^{-1}}\right] = \mathbb{E}_c\left[\|c\|_{A^{-1}}\right]$. Therefore, it holds that

$$\begin{aligned}\underline{R}_{\text{SPO+}}(\Delta) &= -\mathbb{E}_c\left[\frac{2c^T A^{-1}\Delta}{\|c\|_{A^{-1}}}\right] = -\mathbb{E}_c\left[\frac{(c+\phi_0(c))^T A^{-1}\Delta}{\|c\|_{A^{-1}}}\right]\\ &= \mathbb{E}_c\left[\frac{\bar{c}^T A^{-1}c}{\|\bar{c}\|_{A^{-1}}^2}\cdot\frac{\bar{c}^T A^{-1}\Delta}{\|c\|_{A^{-1}}}\right] = \mathbb{E}_c\left[\frac{\bar{c}^T A^{-1}c}{\|c\|_{A^{-1}}}\right]\cdot\frac{-\bar{c}^T A^{-1}\Delta}{\|\bar{c}\|_{A^{-1}}^2}\\ &= \mathbb{E}_c\left[\frac{\bar{c}^T A^{-1}c}{\|c\|_{A^{-1}}}\right]\cdot\frac{\Delta^T A^{-1}\Delta}{\|\bar{c}\|_{A^{-1}}^2},\end{aligned}$$

where the last inequality holds since $(\bar{c}+\Delta)^T A^{-1}\Delta = 0$. Based on the result in Proposition 3.4.2, we have

$$\mathbb{E}_c\left[\frac{\bar{c}^T A^{-1}c}{\|c\|_{A^{-1}}}\right]\geq\mathbb{E}_c\frac{\|c\|_{A^{-1}}^2\min\{\|c\|_{A^{-1}},\|c-\bar{c}\|_{A^{-1}}\}}{\|c-\bar{c}\|_{A^{-1}}^2+\|\bar{c}\|_{A^{-1}}\|c-\bar{c}\|_{A^{-1}}}.$$

Also, let $\epsilon' = C_2(\mu,L,r)\cdot\epsilon$. In the constraint we have

$$\|\bar{c}\|_{A^{-1}} - \frac{\bar{c}^T A^{-1}(\bar{c}+\Delta)}{\|\bar{c}+\Delta\|_{A^{-1}}}\geq\epsilon',$$

and hence $\|\bar{c}\|_{A^{-1}} - \|\bar{c}+\Delta\|_{A^{-1}} \geq \epsilon'$. Since $\|\bar{c}\|_{A^{-1}} \geq \epsilon'$, it holds that $(\|\bar{c}\|_{A^{-1}} - \epsilon')^2 \geq \|\bar{c}+\Delta\|^2_{A^{-1}}$. This implies that $\Delta^T A^{-1}\Delta \geq 2\|\bar{c}\|_{A^{-1}}\epsilon' - \epsilon'^2 \geq \|\bar{c}\|_{A^{-1}}\epsilon' = \|\bar{c}\|_{A^{-1}} C_2(\mu, L, r)\epsilon$. Therefore, we conclude that

$$\delta(\epsilon) \geq \mathbb{E}_c\left[\frac{\min\{\|\bar{c}\|_{A^{-1}}, \|c-\bar{c}\|_{A^{-1}}\}}{\|c-\bar{c}\|^2_{A^{-1}} + \|\bar{c}\|_{A^{-1}}\|c-\bar{c}\|_{A^{-1}}}\right] \cdot \frac{\mu^{9/2}\|\bar{c}\|_{A^{-1}}}{2L^{9/2}} \cdot \epsilon.$$

□

We are now ready to complete the proof of Theorem 3.4.1.

*Proof of Theorem 3.4.1.* From Theorem 3.4.3, we know that

$$\delta(\epsilon; x, \mathbb{P}) \geq \mathbb{E}_{c|x}\left[\frac{\min\{\|\bar{c}\|_{A^{-1}}, \|c-\bar{c}\|_{A^{-1}}\} \cdot \|\bar{c}\|_{A^{-1}}}{\|c-\bar{c}\|^2_{A^{-1}} + \|\bar{c}\|_{A^{-1}}\|c-\bar{c}\|_{A^{-1}}}\right] \cdot \frac{\mu^{9/2}\epsilon}{2L^{9/2}}.$$

Also, by $\frac{\min\{c_1,c_2\}\cdot c_1}{c_2^2+c_1c_2} \geq \frac{c_1^2}{2(c_1^2+c_2^2)}$ for all $c_1, c_2 \neq 0$, we have

$$\delta(\epsilon; x, \mathbb{P}) \geq \mathbb{E}_{c|x}\left[\frac{\|\bar{c}\|^2_{A^{-1}}}{2(\|\bar{c}\|^2_{A^{-1}} + \|c-\bar{c}\|^2_{A^{-1}})}\right] \cdot \frac{\mu^{9/2}\epsilon}{2L^{9/2}}. \tag{3.19}$$

Moreover, for all $\mathbb{P} \in \mathcal{P}_{\alpha,\beta}$, it holds that

$$\begin{aligned}\mathbb{E}_{c|x}\left[\frac{\|\bar{c}\|^2_{A^{-1}}}{\|\bar{c}\|^2_{A^{-1}} + \|c-\bar{c}\|^2_{A^{-1}}}\right] &\geq \mathbb{E}_{c|x}\left[\frac{\|\bar{c}\|^2_{A^{-1}}}{\|\bar{c}\|^2_{A^{-1}} + \|c-\bar{c}\|^2_{A^{-1}}}\middle| \|c-\bar{c}\|_{A^{-1}} \leq \beta \cdot \|\bar{c}\|_{A^{-1}}\right] \\ &\quad \cdot \mathbb{P}_{c|x}(\|c-\bar{c}\|_{A^{-1}} \leq \beta \cdot \|\bar{c}\|_{A^{-1}}) \\ &\geq \frac{\alpha}{1+\beta^2},\end{aligned}$$

and for all $\mathbb{P} \in \mathcal{P}_\beta$, it holds that

$$\mathbb{E}_{c|x}\left[\frac{\|\bar{c}\|^2_{A^{-1}}}{\|\bar{c}\|^2_{A^{-1}} + \|c-\bar{c}\|^2_{A^{-1}}}\right] \geq \frac{\|\bar{c}\|^2_{A^{-1}}}{\|\bar{c}\|^2_{A^{-1}} + \mathbb{E}_{c|x}[\|c-\bar{c}\|^2_{A^{-1}}]} \geq \frac{1}{1+\beta^2}.$$

By applying the above two inequalities to (3.19) we complete the proof. □

## 3.5 Computational Experiments

In this section, we present computational results of synthetic dataset experiments wherein we empirically examine the performance of the SPO+ loss function for training prediction models, using portfolio allocation and cost-sensitive multi-class classification problems as our problem classes. We focus on two classes of prediction models: *(i)* linear models, and *(ii)* two-layer neural networks with 256 neurons in the hidden layer. We compare the performance

of the empirical minimizer of the following four different loss function: *(i)* the previously described SPO loss function (when applicable), *(ii)* the previously described SPO+ loss function, *(iii)* the least squares (squared $\ell_2$) loss function $\ell(\hat{c}, c) = \|\hat{c} - c\|_2^2$, and *(iv)* the absolute ($\ell_1$) loss function $\ell(\hat{c}, c) = \|\hat{c} - c\|_1$. For all loss functions, we use the Adam method of [47] to train the parameters of the prediction models. Note that the loss functions *(iii)* and *(iv)* do not utilize the structure of the feasible region $S$ and can be viewed as purely learning the relationship between cost and feature vectors.

### 3.5.1 Entropy Constrained Portfolio Allocation

First, we consider the portfolio allocation [56] problem with entropy constraint, where the goal is to pick an allocation of assets in order to maximize the expected return while enforcing a certain level of diversity through the use of an entropy constraint (see, e.g., [19]). Alternative formulations of portfolio allocation, including when $S$ is a polyhedron or a polyhedron intersected with an ellipsoid, have been empirically studied in previous works (see, for example [29, 41]). The objective is to minimize $c^T w$ where $c$ is the negative of the expected returns of $d$ different assets, and the feasible region is $S = \{w \in \mathbb{R}^d : w \geq 0, \sum_{i=1}^d w_i = 1, \sum_{i=1}^d w_i \log w_i \leq r\}$ where $r$ is a user-specified threshold of the entropy of portfolio $w$. Note that, due to the differentiability properties of of the optimization oracle $w^*(\cdot)$ in this case (see Lemma 3.5.1 in the Appendix), it is possible to (at least locally) optimize the SPO loss using a gradient method even though SPO loss is not convex.

In our simulations, the relationship between the true cost vector $c$ and its auxiliary feature vector $x$ is given by $c = \phi^{\deg}(Bx) \odot \epsilon$, where $\phi^{\deg}$ is a polynomial kernel mapping of degree deg, $B$ is a fixed weight matrix, and $\epsilon$ is a multiplicative noise term. The features are generated from a standard multivariate normal distribution, we consider $d = 50$ assets, and further details of the synthetic data generation process are provided in the data generation processes and technical details paragraph. To account for the differing distributions of the magnitude of the cost vectors, in order to evaluate the performance of each method we compute a "normalized" SPO loss on the test set. Specifically, let $\hat{g}$ denote a trained prediction model and let $\{\tilde{x}_i, \tilde{c}_i\}_{i=1}^m$ denote the test set, then the normalized SPO loss is defined as $\frac{\sum_{i=1}^m \ell_{\text{SPO}}(\hat{g}(\tilde{x}_i), \tilde{c}_i)}{\sum_{i=1}^m z^*(\tilde{c}_i)}$, where $z^*(\tilde{c}) := \min_{w \in S} \tilde{c}^T w$ is the optimal cost in hindsight. We set the size of the test set to 10000. In all of our experiments, we run 50 independent trials for each setting of parameters.

Figure 3.1 displays the empirical performance of each method. We observe that with a linear hypothesis class, for smaller values of the degree parameters, i.e., deg $\in \{1, 2\}$, all four methods perform comparably, while the SPO and SPO+ methods dominate in cases with larger values of the degree parameters. With a neural net hypothesis class, we observe a similar pattern but, due to the added degree of flexibility, the SPO method dominates the cases with larger values of degree and SPO+ method is the best among all surrogate loss functions. The better results of the $\ell_1$ loss as compared to the squared $\ell_2$ loss might be explained by robustness against outliers. Section 3.5.3 also contains results showing the

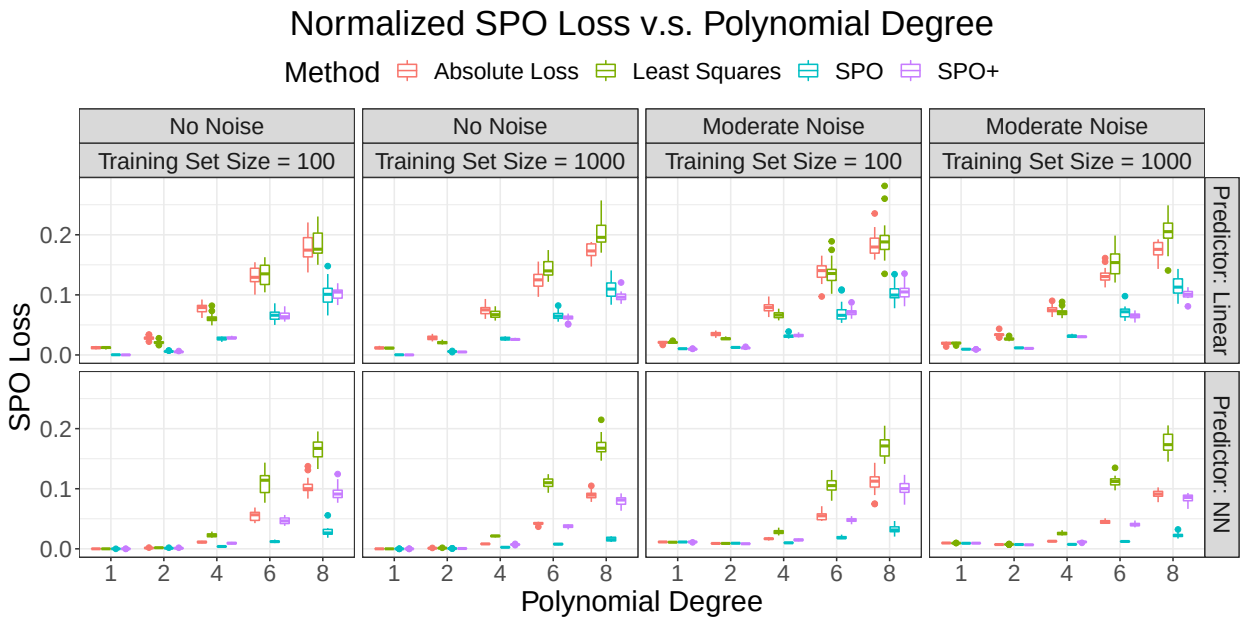



Figure 3.1: Normalized test set SPO loss for the SPO, SPO+, least squares, and absolute loss methods on portfolio allocation instances.

observed convergence of the excess SPO risk, in the case of polynomial degree one, for both this experiment and the cost-sensitive multi-class classification case.

**Data Generation Processes and Technical Details.** Let us describe the process used for generating the synthetic data sets for portfolio allocation instances. In this experiment, we set the number of assets $d = 50$ and the dimension of feature vector $p = 5$. We first generate a weight matrix $B \in \mathbb{R}^{d \times p}$, whereby each entry of $B$ is a Bernoulli random variable with the probability $\mathbb{P}(B_{ij} = 1) = \frac{1}{2}$. We then generate the training data set $\{(x_i, c_i)\}_{i=1}^n$ and the test data set $\{(\tilde{x}_i, \tilde{c}_i)\}_{i=1}^m$ independently according to the following procedure.

1. First we generate the feature vector $x \in \mathbb{R}^p$ from the standard multivariate normal distribution, namely $x \sim \mathcal{N}(0, I_p)$.

2. Then we generate the true cost vector $c \in \mathbb{R}^d$ according to $c_j = \left[1 + \left(1 + \frac{b_j^T x}{\sqrt{p}}\right)^{\deg}\right] \epsilon_j$ for $j = 1, \ldots, d$, where $b_j$ is the $j$-th row of matrix $B$. Here deg is the fixed degree parameter and $\epsilon_j$, the multiplicative noise term, is a random variable which independently generated from the uniform distribution $[1 - \bar{\epsilon}, 1 + \bar{\epsilon}]$ for a fixed noise half width $\bar{\epsilon} \geq 0$. In particular, $\bar{\epsilon}$ is set to 0 for "no noise" instances and 0.5 for "moderate noise" instances.

In Lemma 3.5.1 we show that the optimization oracle $w^*(\cdot)$ is differentiable when the projection of the predicted cost vector $\hat{c}$ is not zero for the entropy constrained portfolio optimization example.

**Lemma 3.5.1.** *Let $T = \{w \in \mathbb{R}^d : w > 0, \mathbf{1}^T w = 1\}$ denote the interior of the probability simplex. For any vector $c \in \mathbb{R}^d$, let $\tilde{c}$ denote the projection of $c$ onto $T$. Let $f(w) = \sum_{i=1}^d -w_i \log(w_i)$ denote the entropy function. For some scalar $r \in (f_{\min}, \underline{\lim}_{w \to \partial T} f(w))$, let $S = \{w \in T : f(w) \leq r\}$. Let $w^*(c) = \arg\min_{w \in S} c^T w$. Then it holds that $w^*(c)$ is differentiable when $\tilde{c} \neq 0$ where $\tilde{c}$ is the projection of $c$ onto the subspace $\{w \in \mathbb{R}^d : \mathbf{1}^T w = 0\}$.*

*Proof.* Let $\text{softmax}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ denote the softmax function, namely

$$\text{softmax}(c) = \left[ \frac{\exp(c_1)}{\sum_{i=1}^d \exp(c_i)}, \ldots, \frac{\exp(c_d)}{\sum_{i=1}^d \exp(c_i)} \right]^T.$$

Using KKT condition, we know that for any $c \in \mathbb{R}^d$ such that $\tilde{c} \neq 0$, there exists some scalar $u(c) \geq 0$ such that $c = -u(c) \cdot \nabla f(w^*(c))$, and therefore $w^*(c) = \text{softmax}(-\tilde{c}/u(c))$. Since the softmax function is differentiable and $\tilde{c}$ is differentiable with respect to $c$, we only need to show that the function $u(c)$ is also differentiable with respect to $c$. Indeed, when $\tilde{c} \neq 0$, we have $f(w^*(c)) = r$, which is equivalent to $f(\text{softmax}(-\tilde{c}/u(c))) = r$. Let $\phi(c, u) = f(\text{softmax}(-\tilde{c}/u))$. Since $\phi(c, u)$ is a decreasing function for $u > 0$, by inverse function theorem we have $\frac{\mathrm{d}u}{\mathrm{d}c} = -\frac{\partial \phi / \partial c}{\partial \phi / \partial u}$, and hence $u(c)$ is also differentiable with respect to $c$. □

### 3.5.2 Cost-Sensitive Multi-Class Classification

Here we consider the cost-sensitive multi-class classification problem. Since this is a multi-class classification problem, the feasible region is simply the unit simplex $S = \{w \in \mathbb{R}^d : w \geq 0, \sum_{i=1}^d w_i = 1\}$ We consider an alternative model for generating the data, whereby the relationship between the true cost vector $c$ and its auxiliary feature vector $x$ is as follows: first we generate a score $s = \sigma(\phi^{\text{deg}}(b^T x) \odot \epsilon)$, where $\phi^{\text{deg}}$ is a degree-deg polynomial kernel, $b$ is a fixed weight vector, $\epsilon$ is a multiplicative noise term, and $\sigma(\cdot)$ is the logistic function. Then the true label is given by $\text{lab} = \lceil 10s \rceil \in \{1, \ldots, 10\}$ and the cost vector $c$ is given by $c_i = |i - \text{lab}|$ for $i = 1, \ldots, 10$. The features are generated from a standard multivariate normal distribution, and further details of the synthetic data generation process are provided in the data generation processes and technical details paragraph. Since the scale of the cost vectors do not change as we change different parameters, we simply compare the test set SPO loss for each method. We still set the size of the test set to 10000 and we run 50 independent trials for each setting of parameters. In addition to the regular SPO+, least squares, and absolute losses, we consider an alternative surrogate loss constructed by considering the SPO+ loss using a log barrier (strongly convex) *approximation* to the unit simplex. That is, we consider the SPO+ surrogate that arises from the set $\tilde{S} := \{w \in \mathbb{R}^d : w \geq 0, \sum_{i=1}^d w_i =$

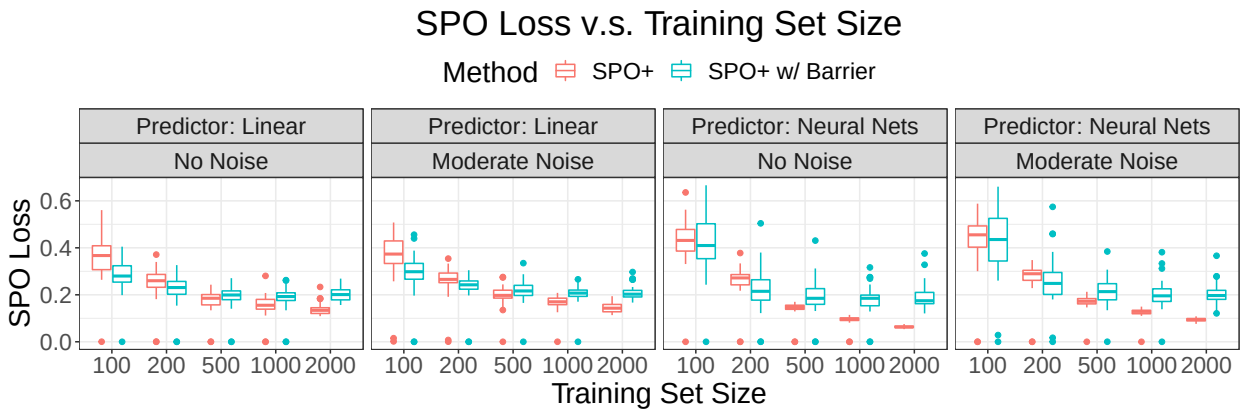



Figure 3.2: Test set SPO loss for the SPO+ methods with different feasible regions on the cost-sensitive multi-class classification instances.

$1, -\sum_{i=1}^{d} \log w_i \leq r\}$ for some $r > 0$. Details about how we chose the value of $r$ are provided in the data generation processes and technical details paragraph.

Herein we focus on the comparison between the standard SPO+ loss and the "SPO+ w/ Barrier" surrogate loss. (The more complete comparison of all the method akin to Figure 3.1 is provided in Figure 3.3.) Figure 3.2 shows a detailed comparison between these alternative SPO+ surrogates as we vary the training set size. Note that the SPO loss is always measured with respect to the standard unit simplex and not the log barrier approximation. Interestingly, we observe that the "SPO+ w/ Barrier" surrogate tends to perform better than the regular SPO+ surrogate when the training set size is small, whereas the regular SPO+ surrogate gradually performs better as the training set size increases. These results suggest that adding a barrier constraint to the feasible region has a type of regularization effect, which may also be explained by our theoretical results. Indeed, adding the barrier constraint makes the feasible region strongly convex, which improves the rate of convergence of the SPO risk. On the other hand, this results in an approximation to the actual feasible region of interest and, eventually for large enough training set sizes, the regularizing benefit of the barrier constraint is outweighed by the cost of this approximation.

**Data Generation Processes and Technical Details.** Let us describe the process used for generating the synthetic data sets for cost-sensitive multi-class classification instances. In this experiment, we set the number of class $d = 10$ and the dimension of feature vector $p = 5$. We first generate a weight vector $b \in \mathbb{R}^p$, whereby each entry of $b$ is a Bernoulli random variable with the probability $\mathbb{P}(b_j = 1) = \frac{1}{2}$. We then generate the training data set $\{(x_i, c_i)\}_{i=1}^{n}$ and the test data set $\{(\tilde{x}_i, \tilde{c}_i)\}_{i=1}^{m}$ independently according to the following procedure.

1. First we generate the feature vector $x \in \mathbb{R}^p$ from the standard multivariate normal

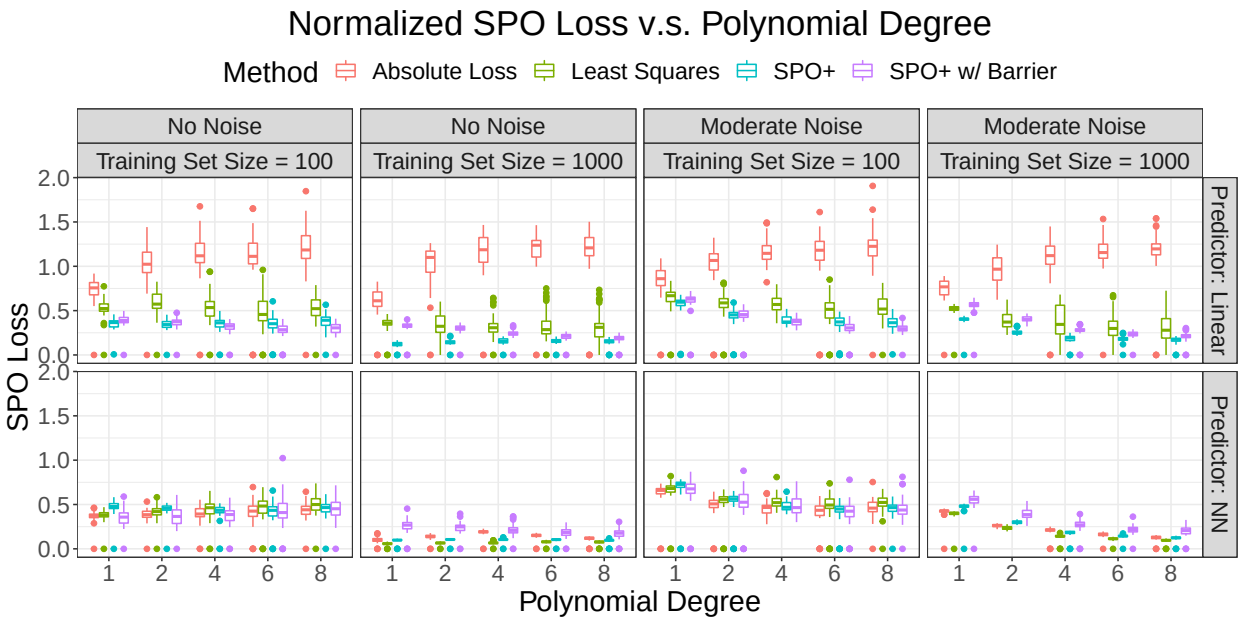



Figure 3.3: Test set SPO loss for the SPO+, least squares, and absolute loss methods on cost-sensitive multi-class classification instances.

distribution, namely $x \sim \mathcal{N}(0, I_p)$.

2. Then we generate the score $s \in (0, 1)$ according to $s = \sigma\left((b^T x)^{\deg} \cdot \text{sign}(b^T x) \cdot \epsilon\right)$, where $\sigma(\cdot)$ is the logistic function. Here $\epsilon$, the multiplicative noise term, is a random variable which independently generated from the uniform distribution $[1 - \bar{\epsilon}, 1 + \bar{\epsilon}]$ for a fixed noise half width $\bar{\epsilon} \geq 0$. In particular, $\bar{\epsilon}$ is set to 0 for "no noise" instances and 0.5 for "moderate noise" instances.

3. Finally we generate the true class label $\text{lab} = \lceil 10s \rceil \in \{1, \ldots, 10\}$ and the true cost vector $c = (c_1, \ldots, c_{10})$ is given by $c_j = |j - \text{lab}|$ for $j = 1, \ldots, 10$.

In the cost-sensitive multi-class classification problem, we consider the SPO+ method using a log barrier approximation to the unit simplex. For the choice of the threshold $r$, according to Assumption 3.4.1 we will need $r > f_{\min}$ and $r < \underline{\lim}_{w \to \partial T} f(w)$. In this log barrier scenario, we have $f_{\min} = d \log d$ and $\underline{\lim}_{w \to \partial T} f(w) = \infty$. Therefore, we pick the threshold $r = 2d \log d$. Of course, one may consider a more careful tuning of this hyper-parameter. Nevertheless, even with our simplistic approach for setting it we observe benefits of the SPO+ loss that uses a log barrier approximation to the unit simplex.

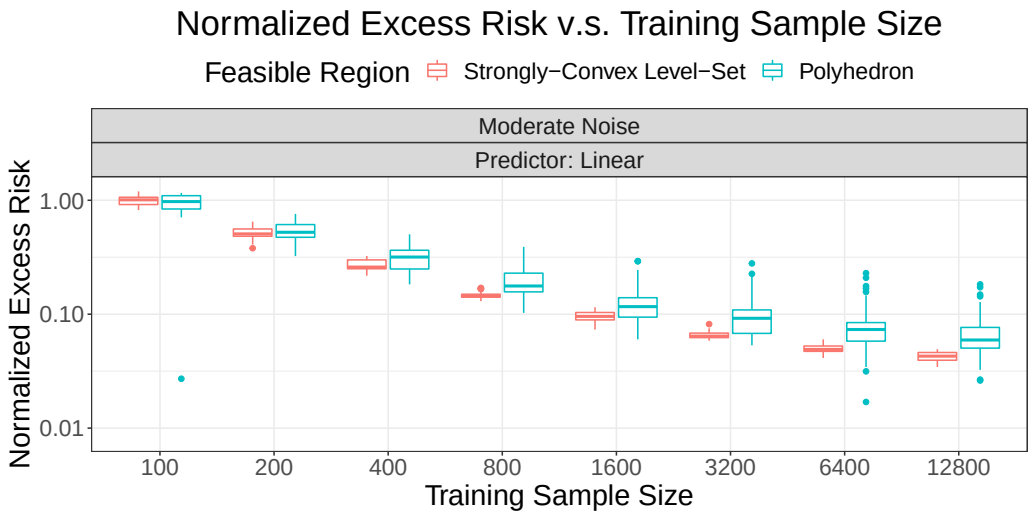



Figure 3.4: Normalized test set excess risk for the SPO+ methods on instances with polyhedron and level-set feasible regions.

### 3.5.3 Excess Risk Comparison

In Figure 3.4, we provide the empirical excess risk comparison of the cases with polyhedral and level-set feasible regions. The case with polyhedral feasible region are the cost-sensitive multi-class classification instances with simplex feasible region, and the case with level-set feasible region are the entropy constrained portfolio optimization problems. The main metric we use in Figure 3.4 is the *normalized excess risk*, which for each case, is defined as the excess risk over the averaged excess risk with sample size $n = 100$. For each type of feasible region, the excess risk is calculated by the difference between the SPO risk of the predictions given by the trained model and the true model. Also, we set polynomial degree equals to one with moderate noises, which means the true model is in the hypothesis class. The main purpose of this plot is not checking if the order of the calibration matches the theoretical results, as these are only worst case guarantees, but qualitatively comparing the convergence of excess risk with different types of feasible regions.

# Chapter 4

# Online Contextual Decision-Making with a Smart Predict-then-Optimize Method

## 4.1 Introduction

Decision-making over time in the presence of uncertainty is a common task across many applications of machine learning. Some typical example problems include online network revenue management, resource allocation, and advertisement bidding. In these settings, there is a trade-off between immediate rewards and rewards received at a later time. This trade-off exists since each decision that is made consumes some of a limited amount of resources. Often, the decision-maker does not have full knowledge of the relevant parameters dictating the amount of the rewards and resources consumed at time $t$, and instead has available contextual information that is related to these parameters and can be used to reduce uncertainty in the decision-making process. Indeed, contextual information such as search history, previous reviews, users characteristics, and many others may be available. For example, in online advertising we may not precisely know the probability that the user would click on a given advertisement, but we may build a machine learning model for predicting this probability based on characteristics of the user and the advertisement.

Recently, there has been a growing interest in the development of machine learning models in the "predict-then-optimize" (or "decision-focused", "end-to-end learning", etc.) setting, where models are trained in a way that is guided by the objectives of a downstream optimization task. See, for example, the works of [20, 25, 29, 44, 30, 40, 68, 48], and the references therein, among others. Prior work in this landscape has primarily been focused on the standard "static" setting where decision-making over time is not a critical aspect and there is no consumption of resources. In this work, we develop a new framework for integrating decision-focused learning methods, using predict-then-optimize losses, into the online decision-making task. Effectively utilizing the structure of the underlying optimization

problem in the decision-making task leads to better decisions more quickly and a better management of the trade-off between immediate and future rewards, which we demonstrate in our numerical experiments. Specifically, we focus on an online contextual stochastic convex optimization problem where we would like to maximize the average (equivalently total) linear reward over time plus a concave utility function that measures the desirability of the average resource consumption levels so far. Our model also includes a convex feasibility constraint on the resource consumption vector. At each time period, the decision-maker is given some contextual features that are associated with the coefficients of the unknown reward objective and resource consumption matrix. We present a "meta-procedure" for online-decision making which involves a prediction step as well as a decision step. In the prediction step, a model is trained for predicting the unknown coefficients based on the history of observed contexts and corresponding coefficients. In the decision step, we use these predictions to solve a linear optimization problem with a known feasible region to make a decision.

Due to the linear structure of the underlying optimization problem in our meta-procedure, we can apply the Smart Predict-then-Optimize (SPO) loss function and its SPO+ convex surrogate loss function developed by [29]. Importantly, the SPO loss function measures the regret of a prediction against the best decision in hindsight and is the ideal loss function to measure the error of the prediction models that we build. Unlike the standard SPO setting, we need to account for the trade-offs present due to the consumption of resources. To do so, we apply the customary technique of introducing dual variables and using primal-dual methods (see, e.g., [4]). As such, at each time period, we update a set of dual variables using the method of online mirror descent [79] and then we update the prediction model by minimizing a surrogate of the SPO loss on a dataset constructed by combining past observations with the current dual variables. A critical part of our contribution involves bridging convergence theory for primal-dual online methods with learning theory in the predict-then-optimize setting. In particular, we prove regret bounds for our overall algorithm that combine the $\mathcal{O}(T^{-1/2})$ convergence of online mirror descent with the convergence of the learning process, the rate of which depends on which surrogate loss function is used. To analyze the latter, we leverage risk bounds and related recent statistical results on the SPO loss and its surrogate loss functions [29, 28, 41, 67, 51]. These results enable us to use a general hypothesis class for fitting the prediction model. More specifically, we are no longer limited to the previously studied linear context or finite policy assumptions [2, 14], and more complex machine learning models, such as random forests and neural networks, may be used. Our bounds hold in both hard and soft resource constraint cases, and we extend prior results using standard upper bound consumption constraints on each resource to arbitrary convex consumption constraints. On the experimental side, we examine the empirical performance of different loss functions in the prediction step of our algorithm. On multi-dimensional knapsack and longest path instances, we observe that the methods which perform best are those that account for both resource consumption via dual variables as well as the structure of the optimization problem via SPO-like loss functions.

Online contextual learning problems have been previously studied under varying assumptions in several different settings. As in our setting, some of these works, including those

that study online linear/convex programming, consider the case where full information is provided after a decision is made (see [54, 6, 4, 42, 50, 16, 49, 83, 52], among others). Other authors have considered bandit and related problems where only partial information is revealed after the decision is made (see [5, 3, 15, 2, 31, 14, 71], among others).

### 4.1.1 Notation

Let $\odot$ represent element-wise multiplication between two vectors. Let $I_p$ denote the $p$ by $p$ identity matrix, and let $e$ denote the vector of all ones in the appropriate dimension. We will make use of a generic given norm $\|\cdot\|$ on $w \in \mathbb{R}^d$, as well as its dual norm $\|\cdot\|_*$ which is defined by $\|c\|_* = \max_{w:\|w\|\leq 1} c^T w$. With a slight abuse of notation, we also let $\|\cdot\|$ refer to a (possibly different) generic given norm on $v \in \mathbb{R}^m$, where which norm we are referring to is clear from the dimension of the corresponding vector. For any convex function $f(\cdot) : F \to \mathbb{R}$ with its domain $F$, let $f^*(\cdot)$ denote its Fenchel conjugate function, namely $f^*(y) := \sup_{x\in F}\{y^T x - f(x)\}$. We also make use of the big $\mathcal{O}$ notation to omit absolute constants.

## 4.2 Online Contextual Convex Optimization and Preliminaries

We now formally describe our online contextual stochastic convex optimization problem, which is prevalent in online decision-making. We assume there are $T$ rounds of decision-making. At each round $t$, we make a decision $w_t \in \mathcal{S} \subseteq \mathbb{R}^d$, and associated with this decision is a "budget consumption vector" $v_t \in \mathbb{R}^m$. Specifically, let $\mathcal{S} \subseteq \mathbb{R}^d$ denote the convex and compact feasible region of the decision variables and let $\mathcal{V} \subseteq \mathbb{R}^m$ denote the closed and convex feasible region of consumption vectors. In addition, there is a "utility function" $u(\cdot) : \mathbb{R}^m \to \mathbb{R}$, assumed to be $L$-Lipschitz and concave with $u(0) = 0$, that describes the consumption vector spending preferences of the decision-maker. We assume that we have full knowledge of $\mathcal{S}$, $\mathcal{V}$, and $u(\cdot)$.

**Example 4.2.1.** *Consider a multi-dimensional knapsack problem where, in each round, the decision-maker receives $d$ different orders and may accept at most $k \leq d$ of them. Each accepted order receives a reward and consumes some of $m$ different resources. The amount of resources available per round is $b \in \mathbb{R}^m$. The selling price vector of any leftover resources is $y \in \mathbb{R}^m$. In this case, the decision space is $\mathcal{S} = \{w \in \mathbb{R}^d : \sum_{j=1}^d w_j \leq k, 0 \leq w \leq e\}$, the resource consumption feasible region is $\mathcal{V} = \{v \in \mathbb{R}^m : v \leq b\}$, and the resource consumption utility function is $u(v) = y^T(b - v)^+$.*

At time $t$, a tuple $(x_t, r_t, V_t)$ is identically and independently drawn from an unknown distribution $\mathbb{P}$, where $r_t \in \mathbb{R}^d$ denotes the reward vector, $V_t \in \mathbb{R}^{d\times m}$ denotes the budget consumption matrix, and $x_t \in \mathbb{R}^p$ denotes the feature vector which contains contextual

information about $r_t$ and $V_t$. The reward vector and consumption matrix are unknown when the decision $w_t \in \mathcal{S}$ needs to be made, while the context vector $x_t$ is given instead. After the decision $w_t$ is made, the actual values of $r_t$ and $V_t$ will be revealed, and we will receive $r_t^T w_t$ as reward and also incur $v_t := V_t^T w_t$ consumption in the budget. Let $r_{\text{avg}}$ and $v_{\text{avg}}$ denote the total averaged reward and consumption values, namely $r_{\text{avg}} := \frac{1}{T}\sum_{t=1}^{T} r_t^T w_t$ and $v_{\text{avg}} := \frac{1}{T}\sum_{t=1}^{T} v_t$. The simultaneous objectives of the decision-maker are: (i) maximize the reward plus the utility of the budget consumption, i.e., $\max\{r_{\text{avg}} + u(v_{\text{avg}})\}$, and (ii) ensure that the average consumption remains feasible, i.e., $v_{\text{avg}} \in \mathcal{V}$.

### 4.2.1 Primal-Dual Formulation and Meta-Procedure

Online decision-making problems, including online linear optimization and bandit problems with constraints, have been well-studied in the machine learning and operations research communities, wherein a common method to address budget consumption utility and/or feasibility constraints is with the primal-dual max-min form of the original problem. In our setting, we will need two sets of dual variables, $\theta$ and $\lambda$, to address both consumption utility and feasibility constraints simultaneously. Let $d_{\mathcal{V}}(\cdot)$ denote the distance function to the set $\mathcal{V}$, measured in the given generic norm $\|\cdot\|$, and let and $\zeta$ be the positive budget penalty parameter. That is, $d_{\mathcal{V}}(\cdot)$ is defined by $d_{\mathcal{V}}(v) := \inf_{\tilde{v}\in\mathcal{V}} \|\tilde{v} - v\|$. Then, for any values of $r_{\text{avg}}$ and $v_{\text{avg}}$ as defined above, we can consider a penalized version of the objective and its primal-dual reformulation using conjugate functions as follows:

$$r_{\text{avg}} + u(v_{\text{avg}}) - \zeta \cdot d_{\mathcal{V}}(v_{\text{avg}}) = \inf_{\lambda\in\Lambda,\theta\in\Theta}\{r_{\text{avg}} - (\lambda^T v_{\text{avg}} - (-u)^*(\lambda)) - \zeta\cdot(\theta^T v_{\text{avg}} - d_{\mathcal{V}}^*(\theta))\}, \quad (4.1)$$

where $\Lambda$ and $\Theta$ are the domains of the conjugate functions $(-u)^*(\cdot)$ and $d_{\mathcal{V}}^*(\cdot)$, respectively. Note that $L$-Lipschitzness of $u(\cdot)$ and 1-Lipschitzness of $d_{\mathcal{V}}(\cdot)$ imply that the domains satisfy $\Lambda \subseteq \{\lambda \in \mathbb{R}^m : \|\lambda\|_* \leq L\}$ and $\Theta \subseteq \{\theta \in \mathbb{R}^m : \|\theta\|_* \leq 1\}$. The main benefit of the introduction of the dual variables is that the primal-dual objective becomes linear in the average reward and consumption whenever the dual variables are fixed. Thus, it is viable to apply an online descent method to the primal-dual min-max problem, which consists of two steps: (i) making a decision by solving an optimization problem with a linear objective, and (ii) updating the dual variables via online descent. Algorithm 4.1 below presents a "meta-procedure" that combines these two steps with a *prediction step* for predicting the reward vector and consumption matrix based on the context $x_t$. We will specify the precise methods for the prediction model and dual variables update later in Section 4.3.

**Algorithm 4.1:** A "meta-procedure" for online contextual decision-making at time $t$

1 Observe feature vector $x_t$;
2 Make predictions $(\hat{r}_t, \hat{V}_t) \leftarrow g_t(x_t)$ for reward and consumption;
3 Make the decision $w_t \leftarrow \arg\max_{w\in\mathcal{S}}\{(\hat{r}_t - \hat{V}_t\lambda_t - \zeta\cdot\hat{V}_t\theta_t)^T w\}$;
4 Observe realized reward $r_t$ and consumption $V_t$;
5 Update dual variables $\theta_{t+1}, \lambda_{t+1}$, and prediction model $g_{t+1}(\cdot)$;

### 4.2.2 Benchmark

In the theoretical part of this work, we compare the performance of the proposed online algorithm against the performance of the optimal static policy, *i.e.*, a policy which knows the distribution $\mathbb{P}$ but only requires to satisfy the resource budget constraints in expectation. The formal definition of the optimal static policy is given below.

**Definition 4.2.1.** *Consider any static policy $\pi(\cdot) : \mathcal{X} \to \mathcal{S}$, and define the expected reward and resource consumption of $\pi(\cdot)$ as*

$$\text{rew}(\pi) := \mathbb{E}_{(x,r,V)\sim\mathbb{P}}[r^T\pi(x)], \text{ and } \text{con}(\pi) := \mathbb{E}_{(x,r,V)\sim\mathbb{P}}[V^T\pi(x)].$$

*Also, define the optimal static reward as the supremum of all feasible static policies, namely*

$$\text{OPT} := \sup_{\pi} \{\text{rew}(\pi) + u(\text{con}(\pi))\}, \quad \text{s.t. con}(\pi) \in \mathcal{V}.$$

Another benchmark would be the optimal adaptive policy, which knows the distribution $\mathbb{P}$ and also takes the history into account. It turns out that the expected total reward of this adaptive policy is upper bounded by the one from the static one [2]. As has been considered in similar settings (see, for example, [24, 13, 2]), we will therefore work with the optimal static policy benchmark defined above.

## 4.3 An Online Algorithm using Predict-then-Optimize and Mirror Descent

In this section, we specify the details for the prediction model and dual variables updates in Algorithm 4.1. In particular, we first describe the predict-then-optimize methodology for learning the prediction model and then describe the online mirror descent method for updating the dual variables.

### 4.3.1 Prediction Model Updating and the SPO Loss

In order to obtain a model for predicting reward vectors and consumption matrices, namely a prediction function $g : \mathbb{R}^p \to \mathbb{R}^d \times \mathbb{R}^{d\times m}$, we may leverage machine learning methods to learn the underlying distribution $\mathbb{P}$ from previously observed data $\{(x_1, r_1, V_1), \ldots, (x_{t-1}, r_{t-1}, V_{t-1})\}$, which are assumed to be independent samples from $\mathbb{P}$. Notice that the optimization subroutine to determine $w_t$ in Algorithm 4.1 involves a *linear* objective function. Ideally, with full knowledge of the distribution $\mathbb{P}$, one would determine $w_t$ by solving the optimization problem

$$\max_{w\in\mathcal{S}} \mathbb{E}_{r,V\sim\mathbb{P}(\cdot|x)} \left[(r - V\lambda - \zeta \cdot V\theta)^T w\right] = \max_{w\in\mathcal{S}} \mathbb{E}_{r,V\sim\mathbb{P}(\cdot|x)}[r - V\lambda - \zeta \cdot V\theta]^T w. \tag{4.2}$$

Due to the linearity of the objective, the above equation implies that it is sufficient to learn the conditional expectation of the "linear cost vector" $c = r - V\lambda - \zeta \cdot V\theta$. Thus $g(x)$ can be thought of as providing estimates of $\mathbb{E}[r|x]$ and $\mathbb{E}[V|x]$, which are then plugged into the corresponding linear optimization problem with feasible region $\mathcal{S}$. This setting is essentially a parametric variant of the the predict-then-optimize framework, where the dual variables $\omega := (\lambda, \theta)$ are parameters that specify a linear cost vector that we would like to learn. In the usual "static" setting without parameters, [29] introduced and studied the SPO loss function, which characterizes the excess cost, or decision error, incurred when making a suboptimal decision due to an imprecise objective cost vector prediction. Let us now adapt the SPO loss to our setting. In the usual case, given a predicted cost vector $\hat{c}$ and a realized cost vector $c$, the SPO loss for a linear optimization problem in maximization format is defined as

$$\ell_{\text{SPO}}(\hat{c}, c) := c^T(w^*(c) - w^*(\hat{c})),$$

where $w^*(\cdot)$ is an optimization oracle for $\mathcal{S}$ satisfying $w^*(c) \in \arg\max_{w \in \mathcal{S}} \left\{c^T w\right\}$. In our setting, we need to consider a parametric variant of the SPO loss where the dual variables are parameters affecting the objective cost vectors. In particular, given a prediction $\hat{\mu} := (\hat{r}, \hat{V})$, realization $\mu := (r, V)$, dual variables $\omega := (\lambda, \theta)$, as well as fixed budget penalty parameter $\zeta > 0$, the SPO loss of the optimization problem (4.2) is defined as

$$\ell_{\text{SPO}}(\hat{\mu}, \mu; \omega) := (r - V\lambda - \zeta \cdot V\theta)^T(w^*(\mu; \omega) - w^*(\hat{\mu}; \omega)),$$

where $w^*(\cdot)$ denotes the optimization oracle, which is defined as a function satisfying

$$w^*(\mu; \omega) \in \arg\max_{w \in \mathcal{S}} \left\{(r - V\lambda - \zeta \cdot V\theta)^T w\right\}, \text{ for all } \mu \in \mathbb{R}^d \times \mathbb{R}^{d \times m} \text{ and } \omega \in \Lambda \times \Theta.$$

Since the SPO loss function is usually non-convex and even possibly non-continuous, several surrogate loss functions have been introduced. For example, [29] introduce the SPO+ loss function that accounts for the structure of $\mathcal{S}$ when training the prediction model. This loss function is defined as

$$\ell_{\text{SPO+}}(\hat{c}, c) := \max_{w \in \mathcal{S}} \{(c - 2\hat{c})^T w\} + 2\hat{c}^T w^*(c) - c^T w^*(c).$$

On the other hand, more standard prediction error loss functions, like the squared $\ell_2$ loss of the linear cost vector, may be considered. Let $\ell(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be any surrogate loss function of the standard SPO loss, which takes cost vector inputs, including possibly itself. Then, just as we defined an extension of the SPO loss to our setting with dual variables, we can also extend the surrogate loss $\ell$ by defining

$$\ell(\hat{\mu}, \mu; \omega) := \ell(\hat{r} - \hat{V}\lambda - \zeta \cdot \hat{V}\theta, r - V\lambda - \zeta \cdot V\theta).$$

Given a surrogate loss function, we use empirical risk minimization to update the prediction model $g_t$ at each step of our online decision-making method. Specifically, let $\mathcal{H}$ refer to a hypothesis class of predictor functions mapping features $x$ to predictors $(\hat{r}, \hat{V})$ of the reward

vector and resource consumption matrix. Then, the prediction model used at iteration $t$ is chosen by

$$g_t \leftarrow \arg\min_{g\in\mathcal{H}} \sum_{s=1}^{t-1} \ell(g(x_s), \mu_s; \omega_t).$$

It is worthwhile to notice that the dual variables used in the loss function are those of the current iteration instead of the previous ones. Intuitively, the current set of dual variables $\omega_t$ is closer to the optimal dual variables of the offline expected problem and hence it leads to a better prediction model.

A desirable property of the surrogate loss $\ell$ is that the empirical risk minimizer for $\ell$ has small excess risk with respect to the SPO loss. This property is formalized below in Assumption 4.3.1, wherein we measure the excess risk by comparing with the ground truth conditional expectation function of the reward vector and resource consumption matrix, namely $g^*(x) := \mathbb{E}_{\mu\sim\mathbb{P}(\cdot|x)}[\mu]$. Let us also define the expected risk functions for the two losses by $R_{\text{SPO}}(g;\omega) := \mathbb{E}_{(x,\mu)\sim\mathbb{P}}[\ell_{\text{SPO}}(g(x),\mu;\omega)]$ and $R_\ell(g;\omega) := \mathbb{E}_{(x,\mu)\sim\mathbb{P}}[\ell(g(x),\mu;\omega)]$.

**Assumption 4.3.1.** *There exist constants $\kappa_{\text{risk}}, \alpha > 0$ such that, for any integer $n > 0$ and uniformly over all dual variables $\omega \in \Lambda \times \Theta$, the empirical surrogate loss optimal predictor*

$$\hat{g}^n := \arg\min_{g\in\mathcal{H}} \left\{ \sum_{i=1}^{n} \ell(g(x_i), \mu_i; \omega) \right\},$$

*satisfies the following excess true SPO risk guarantee*

$$\mathbb{E}[R_{\text{SPO}}(\hat{g}^n;\omega)] - R_{\text{SPO}}(g^*;\omega) \le \kappa_{\text{risk}} \cdot n^{-\alpha},$$

*where the expectation is taken with respect to i.i.d. samples $\{(x_i,\mu_i)\}_{i=1}^n$ drawn from $\mathbb{P}$.*

In general, the rate of convergence $\alpha$ and the constant $\kappa_{\text{risk}}$ in Assumption 4.3.1 depend on the properties of the surrogate loss function, the decision feasible region $\mathcal{S}$, the underlying distribution $\mathbb{P}$, and the complexity of the hypothesis class $\mathcal{H}$. Assumption 4.3.1 is closely tied to uniform calibration properties of the surrogate loss $\ell$ with respect to the SPO loss. In fact, the following remark demonstrates that uniform calibration and an excess risk bound for $\ell$ are sufficient conditions for Assumption 4.3.1.

**Remark 4.3.1.** *Suppose that the the surrogate loss function $\ell(\cdot,\cdot)$ is uniformly calibrated with respect to the true SPO loss in the standard setting. Namely, for some constants $\kappa_1$ and $\beta$, for any distribution $\tilde{\mathbb{P}}$ over cost vectors $c$, we have*

$$\mathbb{E}_{c\sim\tilde{\mathbb{P}}}[\ell(\hat{c},c) - \ell(\mathbb{E}[c],c)] \le \kappa_1 \cdot \epsilon^\beta \Rightarrow \mathbb{E}_{c\sim\tilde{\mathbb{P}}}[\ell_{\text{SPO}}(\hat{c},c) - \ell_{\text{SPO}}(\mathbb{E}[c],c)] \le \epsilon,$$

*for all $\hat{c} \in \mathbb{R}^d$ and $\epsilon > 0$. Suppose further that the empirical surrogate loss optimal predictor has an excess risk bound that holds uniformly over all dual variables $\omega \in \Lambda \times \Theta$, i.e., for some constants $\kappa_2$ and $\gamma$ we have*

$$\mathbb{E}[R_\ell(\hat{g}^n;\omega)] - R_\ell(g^*;\omega) \le \kappa_2 \cdot n^{-\gamma},$$

*for any $n > 0$ and where the expectation is taken with respect to i.i.d. samples $\{(x_i, \mu_i)\}_{i=1}^n$ drawn from $\mathbb{P}$. Then, under both of these conditions, it holds that*

$$\mathbb{E}[R_{\text{SPO}}(\hat{g}^n; \omega)] - R_{\text{SPO}}(g^*; \omega) \leq (\kappa_2/\kappa_1)^{1/\beta} \cdot n^{-\gamma/\beta}.$$

We note that the two conditions in Remark 4.3.1 often hold for many choices of surrogate loss $\ell$ and hypothesis classes $\mathcal{H}$. Indeed, recent works including [67, 41, 51] have examined sufficient conditions under which uniform calibration holds for the SPO loss. Some conditions require an additional restriction on the class of distributions $\bar{\mathbb{P}}$ over cost vectors, but such restrictions will often be satisfied in practice in our setting. Furthermore, for most common choices of surrogate losses $\ell$, e.g., convex and Lipschitz losses like squared $\ell_2$ and SPO+, the required excess risk bound will hold. Indeed, for most common surrogates, due to the boundedness of the dual variable domains $\Lambda$ and $\Theta$, boundedness of the Rademacher complexity of $\mathcal{H}$ would be a sufficient condition to ensure that the bound holds uniformly over the dual variables.

**Lemma 4.3.1.** *Suppose Assumption 4.3.1 holds. For any policy $\pi(\cdot) : \mathcal{X} \to \mathcal{S}$ and $T \geq 1$, Algorithm 4.2 satisfies*

$$\mathcal{R}_g(T) := \mathbb{E}\left[\sum_{t=1}^{T}(r_t - V_t\lambda_t - \zeta \cdot V_t\theta_t)^T(\pi(x_t) - w_t)\right] \leq \kappa_{\text{risk}} \cdot \mathcal{O}(T^{1-\alpha}).$$

*Proof.* Let $\bar{w}_t := w^*(g^*(x_t); \omega_t)$, where recall that $g^*(x) := \mathbb{E}_{\mu\sim\mathbb{P}(\cdot|x)}[\mu]$ is the Bayes estimator (i.e., the ground truth model). Since Assumption 4.3.1 holds (in particular uniformly over $\omega \in \Lambda \times \Theta$), for any $t \in \{1, \ldots, T\}$, we have

$$\begin{aligned}
&\mathbb{E}_{(x_1,\mu_1),\ldots,(x_{t-1},\mu_{t-1})\sim\mathbb{P}^{t-1}}[R_{\text{SPO}}(g_t; \omega_t) - R_{\text{SPO}}(g^*; \omega_t)] \\
=&\,\mathbb{E}_{(x_1,\mu_1),\ldots,(x_{t-1},\mu_{t-1})\sim\mathbb{P}^{t-1}}\left[\mathbb{E}_{(x_t,\mu_t)\sim\mathbb{P}}[(r_t - V_t(\lambda_t + \zeta \cdot \theta_t))^T(\bar{w}_t - w_t) \mid \mathcal{F}_{t-1}]\right] \\
=&\,\mathbb{E}_{(x_1,\mu_1),\ldots,(x_t,\mu_t)\sim\mathbb{P}^t}\left[(r_t - V_t(\lambda_t + \zeta \cdot \theta_t))^T(\bar{w}_t - w_t)\right] \leq \kappa_{\text{risk}} \cdot (t-1)^{-\alpha}.
\end{aligned}$$

Hence, it holds that

$$\mathbb{E}\left[(r_t - V_t(\lambda_t + \zeta \cdot \theta_t))^T(\bar{w}_t - w_t)\right] \leq \kappa_{\text{risk}} \cdot (t-1)^{-\alpha} \tag{4.3}$$

where the expectation above is with respect to all randomness of Algorithm 4.2. Also, since $g^*(x_t)$ is the Bayes estimator, for any policy $\pi$, it holds that

$$\mathbb{E}_{(x_t,\mu_t)\sim\mathbb{P}}[(r_t - V_t(\lambda_t + \zeta \cdot \theta_t))^T(\pi(x_t) - \bar{w}_t) \mid \mathcal{F}_{t-1}] \leq 0,$$

and

$$\mathbb{E}[(r_t - V_t(\lambda_t + \zeta \cdot \theta_t))^T(\pi(x_t) - \bar{w}_t)] \leq 0.$$

Therefore, combining (4.3) with the above yields

$$\mathbb{E}_{(x_t,\mu_t)\sim\mathbb{P}}[(r_t - V_t(\lambda_t + \zeta \cdot \theta_t))^T(\pi(x_t) - w_t)] \leq \kappa_{\text{risk}} \cdot (t-1)^{-\alpha}.$$

Then by taking the summation over $t = 1, \ldots, T$, we have

$$\mathcal{R}_g(T) \leq \kappa_{\text{risk}} \cdot \mathcal{O}(T^{1-\alpha}).$$

□

### 4.3.2 Dual Variable Updates with Online Mirror Descent

In this section, we describe how we use online mirror descent method to update the dual variables $\lambda_{t+1}$ and $\theta_{t+1}$. Based on the reformulation (4.1), we define convex functions

$$\xi_t(\lambda) := -v_t^T\lambda + (-u)^*(\lambda) \text{ and } \phi_t(\theta) := -v_t^T\theta + d_{\mathcal{V}}^*(\theta),$$

where we recall that $v_t = V_t^T w_t$. Note that the domains of $\xi_t(\cdot)$ and $\phi_t(\cdot)$ are the same as the domains of the corresponding conjugate functions, i.e., the previously defined sets $\Lambda$ and $\Theta$, respectively. Let $h_\Lambda(\cdot), h_\Theta(\cdot)$ be differentiable and 1-strongly convex (with respect to the dual norm of the norm on $v \in \mathbb{R}^m$) functions on $\Lambda, \Theta$, respectively, and let $B_{h_\Lambda}(\cdot,\cdot), B_{h_\Theta}(\cdot,\cdot)$ denote their respective Bregman distances. For example, $B_{h_\Lambda}(\cdot)$ is defined by

$$B_{h_\Lambda}(\lambda_1, \lambda_2) := h_\Lambda(\lambda_1) - h_\Lambda(\lambda_2) - \nabla h_\Lambda(\lambda_2)^T(\lambda_1 - \lambda_2).$$

Then, online mirror descent uses the following update schemes for the dual variables:

$$\lambda_{t+1} \leftarrow \arg\min_{\lambda\in\Lambda}\{\eta_\lambda \nabla\xi_t(\lambda_t)^T\lambda + B_{h_\Lambda}(\lambda, \lambda_t)\}, \quad \theta_{t+1} \leftarrow \arg\min_{\theta\in\Theta}\{\eta_\theta \nabla\phi_t(\theta_t)^T\theta + B_{h_\Theta}(\theta, \theta_t)\},$$

where $\eta_\lambda, \eta_\theta > 0$ are the "step-size" parameters. Here we abuse notation slightly and let $\nabla$ refer to any subgradient of the functions $\xi_t(\cdot)$ and $\phi_t(\cdot)$. We assume that we can efficiently calculate such subgradients, i.e., by evaluating the subproblems defining the conjugate functions. We also need to be able to efficiently calculate the solution of the above subproblems, which depends on the structure of the Bregman distances. For example, recall that the online mirror descent method is the same as the online projected gradient descent method when the norm and Bregman distances are Euclidean. The following lemma provides an upper bound of the regret from the online mirror descent method.

**Lemma 4.3.2.** *[Theorem 2.15 in [79]] Let $D_\Lambda, D_\Theta$ be upper bounds on the Bregman distances so that $B_{h_\Lambda}(\lambda_1, \lambda_2) \le D_\Lambda$ and $B_{h_\Theta}(\theta_1, \theta_2) \le D_\Theta$ for all $\lambda_1, \lambda_2 \in \Lambda$ and $\theta_1, \theta_2 \in \Theta$. Let $G_\Lambda, G_\Theta$ be upper bounds on norms of the subgradients so that $\|\nabla\xi_t(\lambda_t)\| \le G_\Lambda$ and $\|\nabla\phi_t(\theta_t)\| \le G_\Theta$ for all $t = 1, \ldots, T$. If we use the constant step-sizes $\eta_\lambda \leftarrow \frac{D_\Lambda}{G_\Lambda\sqrt{T}}$ and $\eta_\theta \leftarrow \frac{D_\Theta}{G_\Theta\sqrt{T}}$, then for all $\lambda \in \Lambda$ and all $\theta \in \Theta$ it holds that*

$$\mathcal{R}_\lambda(T) := \sum_{t=1}^{T}(\xi_t(\lambda_t) - \xi_t(\lambda)) \le 2G_\Lambda\sqrt{D_\Lambda T}, \quad \mathcal{R}_\theta(T) := \sum_{t=1}^{T}(\phi_t(\theta_t) - \phi_t(\theta)) \le 2G_\Theta\sqrt{D_\Theta T}.$$

Note that, due to the bounds on the norm of dual variables in the domains $\Lambda$ and $\Theta$, the Bregman constants usually satisfy that $\sqrt{D_\Lambda} = \mathcal{O}(L)$ (recall that $L$ is the Lipschitz constant for $u(\cdot)$), and $\sqrt{D_\Theta} = \mathcal{O}(1)$. Indeed, in the Euclidean case this is guaranteed to be true. Now we have specified the update rules for both the prediction model and the dual variables, we can fully state an implmentable version of the meta algorithm presented previously. This implementable algorithm to solve our online decision-making problem is presented in Algorithm 4.2.

**Algorithm 4.2:** An implementable algorithm for online contextual decision-making

**input :** Budget penalty parameter $\zeta$ and surrogate loss function $\ell(\cdot,\cdot;\cdot)$.

Initialize dual variables $\theta_1, \lambda_1$, and prediction model $g_1(\cdot)$;

**for** $t = 1, \ldots, T$ **do**

- Observe feature vector $x_t$;
- Make predictions $(\hat{r}_t, \hat{V}_t) \leftarrow g_t(x_t)$ for reward and consumption;
- Make the decision $w_t \leftarrow \arg\max_{w\in\mathcal{S}}\{(\hat{r}_t - \hat{V}_t\lambda_t - \zeta\cdot\hat{V}_t\theta_t)^T w\}$;
- Observe realized reward $r_t$ and consumption $V_t$;
- Update dual variable $\lambda_{t+1} \leftarrow \arg\min_{\lambda\in\Lambda}\left\{\eta_\lambda \nabla\xi_t(\lambda_t)^T\lambda + B_{h_\Lambda}(\lambda,\lambda_t)\right\}$;
- Update dual variable $\theta_{t+1} \leftarrow \arg\min_{\theta\in\Theta}\left\{\eta_\theta \nabla\phi_t(\theta_t)^T\theta + B_{h_\Theta}(\theta,\theta_t)\right\}$;
- Update prediction model $g_{t+1} \leftarrow \arg\min_{g\in\mathcal{H}}\{\sum_{s=1}^{t}\ell(g(x_s),\mu_s;\omega_{t+1})\}$;

## 4.4 Regret Bounds and Analysis

In this section, we present the regret analysis of Algorithm 4.2 in two cases: hard and soft constraints.

### 4.4.1 Hard Constraints

We assume the starting point of the budget consumption, which we naturally assume to be the zero vector without loss of generality, is inside the consumption feasible region $\mathcal{V}$. The hard constraints case is when we add a stopping condition to Algorithm 4.2 that terminates whenever the current resource consumption vector violates the constraints enforced by $\mathcal{V}$, i.e., whenever it leaves the feasible region $\mathcal{V}$. We introduce a stopping time $\tau$ that is the first time before time $T$ that the constraints are violated, i.e., $\tau := \min\{t \le T : \frac{1}{T}\sum_{s=1}^{t} V_s^T w_s \notin \mathcal{V}\}$. Most previous works with hard resource constraints, for example, [2] and [49], consider the case of an upper bound budget constraint for each resource, i.e., $\mathcal{V} = \{v : v \le b\}$ for some $b \in \mathbb{R}^m$. In such cases, the online algorithm will terminate immediately whenever it violates any of the resource constraints. In contrast, let $B_\mathcal{V}$ denote the distance from zero to the boundary of the generic closed and convex set $\mathcal{V}$. The constant $B_\mathcal{V}$ will be important in our analysis as it demonstrates how the structure of $\mathcal{V}$ affects the constant in the regret bound. In fact, $B_\mathcal{V}$ precisely generalizes constants appearing in previous works considering only upper bound constraints, for example, [2]. Indeed, when $\mathcal{V} = \{v : v \le b\}$ for some $b \in \mathbb{R}^m$ with each component positive, then it holds that $B_\mathcal{V} = \min_{i=1,\ldots,m} b_i > 0$. The following lemma provides an important property of the constant $B_\mathcal{V}$.

**Lemma 4.4.1.** *For any $v \notin \mathcal{V}$ and $\kappa \in [0,1]$, there exists $\theta \in \Theta$ such that*

$$\kappa \cdot d^*_\mathcal{V}(\theta) - \theta^T v \le (\kappa - 1)\cdot B_\mathcal{V}.$$

*Proof.* Since $v \notin \mathcal{V}$ and $\mathcal{V}$ is a closed convex set, by the separating hyperplane theorem, there exists a vector $\theta \in \mathbb{R}^m$ with $\|\theta\|_* = 1$ such that for any $v^\dagger \in \mathcal{V}$, it holds that $\theta^T v^\dagger < \theta^T v$. Let $\tilde{v}' \in \mathbb{R}^m$ be such that $\|\tilde{v}'\| = 1$ and $\theta^T \tilde{v}' = 1$. Since $0 \in \mathcal{V}$ and $\sup_{v^\dagger \in \mathcal{V}}\{\theta^T v^\dagger\} < \theta^T v < +\infty$, there exists a constant $\iota \geq 0$ such that $v' \leftarrow \iota \cdot \tilde{v}' \in \partial\mathcal{V}$. Therefore, it holds that $d^*_{\mathcal{V}}(\theta) \geq \theta^T v' - d_{\mathcal{V}}(v') = \theta^T v' = \|v'\| \geq B_{\mathcal{V}}$.

Now consider an arbitrary $\tilde{v}^\circ \in \mathbb{R}^m$ and, by the definition of the distance function, let $v^\circ \in \mathcal{V}$ be such that $\|\tilde{v}^\circ - v^\circ\| = d_{\mathcal{V}}(\tilde{v}^\circ)$. Then, it holds that

$$d^*_{\mathcal{V}}(\theta) \geq \theta^T v^\circ = \theta^T \tilde{v}^\circ + \theta^T(v^\circ - \tilde{v}^\circ) \geq \theta^T \tilde{v}^\circ - \|v^\circ - \tilde{v}^\circ\| = \theta^T \tilde{v}^\circ - d_{\mathcal{V}}(\tilde{v}^\circ).$$

Since the above is true for arbitrary $\tilde{v}^\circ$, by taking the supremum over the closed set $\mathcal{V}$ and using the fact that $\sup_{v^\dagger \in \mathcal{V}}\{\theta^T v^\dagger\} < \theta^T v < +\infty$, we have that $d^*_{\mathcal{V}}(\theta) = \theta^T v^\circ$ for some $v^\circ \in \mathcal{V}$. Then it holds that

$$\begin{aligned}\kappa \cdot d^*_{\mathcal{V}}(\theta) - \theta^T v &\leq \kappa \cdot d^*_{\mathcal{V}}(\theta) - \theta^T v^\circ \\ &= (\kappa - 1) \cdot d^*_{\mathcal{V}}(\theta) \\ &\leq (\kappa - 1) \cdot B_{\mathcal{V}},\end{aligned}$$

where the last inequality holds since $\kappa - 1 \leq 0$. □

We make the following boundedness assumption on the distribution.

**Assumption 4.4.1.** *Suppose there exists a constant $D_v \geq 1$, such that for any $w \in \mathcal{S}$, it holds that $\|V^T w\| \leq D_v$ with probability 1. Let the constant $\kappa_{\text{MD}} := D_v \cdot (\zeta\sqrt{D_\Theta} + \sqrt{D_\Lambda})$.*

In the hard constraints case, let $r^\tau_{\text{avg}}$ and $v^\tau_{\text{avg}}$ denote the total averaged reward and consumption with stopping time $\tau$, namely $r^\tau_{\text{avg}} := \frac{1}{\tau}\sum_{t=1}^\tau r_t^T w_t$ and $v^\tau_{\text{avg}} := \frac{1}{\tau}\sum_{t=1}^\tau v_t$. Below we provide our main theorem in the hard constraint case, which provides the regret bound of Algorithm 4.2.

**Theorem 4.4.1.** *Suppose that Assumptions 4.3.1 and 4.4.1 hold, and that the budget penalty parameter $\zeta$ satisfies $\zeta \geq \frac{\text{OPT}}{B_{\mathcal{V}}}$. Then Algorithm 4.2 has the following guarantee:*

$$\text{OPT} - \mathbb{E}[r^\tau_{\text{avg}} + u(v^\tau_{\text{avg}})] \leq \kappa_{\text{MD}} \cdot \mathcal{O}(T^{-1/2}) + \kappa_{\text{risk}} \cdot \mathcal{O}(T^{-\alpha}).$$

We remark that the constant $\kappa_{\text{MD}}$ will usually satisfy $\kappa_{\text{MD}} = D_v \cdot (\zeta\mathcal{O}(1) + \mathcal{O}(L))$ for most choices of Bregman functions. In general, given the required lower bound of $\zeta$ in Theorem 4.4.1, the best value of the constant $\kappa_{\text{MD}}$ will be $\mathcal{O}(D_v \cdot (\frac{\text{OPT}}{B_{\mathcal{V}}}\sqrt{D_\Theta} + \sqrt{D_\Lambda}))$ whenever we are able to set $\zeta = \mathcal{O}(\frac{\text{OPT}}{B_{\mathcal{V}}})$. The dependence on the term $\frac{\text{OPT}}{B_{\mathcal{V}}}$ in the regret bound is natural, since, if the budget starting point is very close to the boundary of the feasible set (or equivalently, in the budget upper bound case, one of the resource budget values is very close to zero), then Algorithm 4.2 is likely to terminate in the first several iterations leading to a poor regret bound.

For convenience in our proofs, let us introduce some new notation. Let $\mathcal{F}_{t-1}$ denote the $\sigma$-field of information revealed up to the start of iteration $t$, i.e., the $\sigma$ field of $\{(x_1, \mu_1), \ldots, (x_{t-1}, \mu_{t-1})\}$. Let $\overline{\mathcal{R}}_\lambda(T) := 2G_\Lambda\sqrt{D_\Lambda T}$ and $\overline{\mathcal{R}}_\theta(T) := G_\Theta\sqrt{D_\Theta T}$ denote the upper bounds of the regret of online mirror descent from Lemma 4.3.2. The following lemma presents the regret from the suboptimality of the dual variables in the hard constraint case.

**Lemma 4.4.2.** *For any feasible policy $\pi(\cdot) : \mathcal{X} \to \mathcal{S}$ and $T \geq 1$, Algorithm 4.2 satisfies*

$$(A): \quad \mathbb{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{\tau}(V_t\theta_t)^T(\pi(x_t)-w_t)\right] \leq \left(\frac{\tau}{T}-1\right)B_{\mathcal{V}} + \frac{\overline{\mathcal{R}}_\theta(T)}{T}, \text{ and}$$

$$(B): \quad \mathbb{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{\tau}(V_t\lambda_t)^T(\pi(x_t)-w_t)\right] \leq \frac{\tau}{T}\cdot(-u)(\mathrm{con}(\pi)) - \mathbb{E}[(-u)(v^\tau_{\mathrm{avg}})] + \frac{\overline{\mathcal{R}}_\lambda(T)}{T}.$$

*Proof.* Let us first prove inequality $(A)$. Since $\pi$ is a feasible policy, the Fenchel-Young inequality yields

$$\mathbb{E}\left[\pi(x_t)^T V_t\theta_t - d^*_{\mathcal{V}}(\theta_t)|\mathcal{F}_{t-1}\right] = (\mathrm{con}(\pi))^T\theta_t - d^*_{\mathcal{V}}(\theta_t) \leq d_{\mathcal{V}}(\mathrm{con}(\pi)) = 0. \tag{4.4}$$

Also, Lemma 4.3.2 guarantees that for any $\theta \in \Theta$, we have

$$\sum_{t=1}^{\tau}(\phi_t(\theta_t) - \phi_t(\theta)) \leq \overline{\mathcal{R}}_\theta(T).$$

Given the definition of $\phi(\cdot)$, which is $\phi_t(\theta') = -v_t^T\theta' + d^*_{\mathcal{V}}(\theta')$, and $v_t = V_t^T w_t$, the above is equivalent to

$$\sum_{t=1}^{\tau}(-w_t^T V_t\theta_t + d^*_{\mathcal{V}}(\theta_t) + w_t^T V_t\theta - d^*_{\mathcal{V}}(\theta)) \leq \overline{\mathcal{R}}_\theta(T) \tag{4.5}$$

Therefore, for any $\theta \in \Theta$, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{\tau}(V_t\theta_t)^T(\pi(x_t)-w_t)\right] \\
\leq&\,\mathbb{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{\tau}\left(\pi(x_t)^T V_t\theta_t - w_t^T V_t\theta + d^*_{\mathcal{V}}(\theta) - d^*_{\mathcal{V}}(\theta_t)\right)\right] + \frac{\overline{\mathcal{R}}_\theta(T)}{T} \\
\leq&\,\mathbb{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{\tau}\left(-w_t^T V_t\theta + d^*_{\mathcal{V}}(\theta)\right)\right] + \frac{\overline{\mathcal{R}}_\theta(T)}{T} \\
=&\,\mathbb{E}\left[\frac{\tau}{T}\cdot d^*_{\mathcal{V}}(\theta) - \theta^T v^\tau_{\mathrm{avg}}\right] + \frac{\overline{\mathcal{R}}_\theta(T)}{T}.
\end{aligned}$$

where the first inequality comes from (4.5), and the second inequality comes from (4.4).

If $\tau = T$, we pick $\theta \leftarrow 0$, and it holds that $\frac{\tau}{T} \cdot d^*_{\mathcal{V}}(\theta) = \theta^T v^\tau_{\text{avg}} = 0$, which implies $(A)$. If $\tau < T$, it implies $v^\tau_{\text{avg}} \notin \mathcal{V}$. Then following the results in Lemma 4.4.1 we know that there exists $\theta \in \Theta$ such that

$$\frac{\tau}{T} \cdot d^*_{\mathcal{V}}(\theta) - \theta^T v^\tau_{\text{avg}} \leq \left(\frac{\tau}{T} - 1\right) \cdot B_{\mathcal{V}}.$$

Therefore, for both $\tau = T$ and $\tau < T$, we have $(A)$.

Let us then prove inequality $(B)$. First, the Fenchel-Young inequality again yields

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{\tau} \mathbb{E}\left[\pi(x_t)^T V_t \lambda_t - (-u)^*(\lambda_t) | \mathcal{F}_{t-1}\right] &= \frac{1}{T}\sum_{t=1}^{\tau} \left(\text{con}(\pi)^T \lambda_t - (-u)^*(\lambda_t)\right) \\ &\leq \frac{1}{T}\sum_{t=1}^{\tau} (-u)(\text{con}(\pi)) = \frac{\tau}{T}(-u)(\text{con}(\pi)). \end{aligned} \tag{4.6}$$

Recall the definition of $\phi(\cdot)$, which is $\phi(\lambda) = -w_t^T V_t \lambda + (-u)^*(\lambda)$. Then, by following the results in Lemma 4.3.2, for any $\lambda \in \Lambda$, it holds that

$$\sum_{t=1}^{\tau} \left(w_t^T V_t \lambda - (-u)^*(\lambda) - w_t^T V_t \lambda_t + (-u)^*(\lambda_t)\right) = \sum_{t=1}^{\tau} (\xi_t(\lambda_t) - \xi_t(\lambda)) \leq \overline{\mathcal{R}}_\lambda(T). \tag{4.7}$$

Let $v' \leftarrow \frac{T}{\tau} \cdot v^\tau_{\text{avg}}$, and pick $\lambda \in \Lambda$ such that $(-u)(v') = (v')^T \lambda - (-u)^*(\lambda)$. Then, using concavity of $u(\cdot)$ and $u(0) = 0$, it holds that

$$\begin{aligned} (-u)(v^\tau_{\text{avg}}) &= (-u)\left(\frac{\tau}{T} \cdot v' + \left(1 - \frac{\tau}{T}\right) \cdot 0\right) \\ &\leq \frac{\tau}{T} \cdot (-u)(v') + \left(1 - \frac{\tau}{T}\right) \cdot (-u)(0) \\ &= \frac{\tau}{T} \cdot (\lambda^T v' - (-u)^*(\lambda)) \\ &= \frac{1}{T} \cdot \sum_{t=1}^{\tau} (w_t^T V_t \lambda - (-u)^*(\lambda)). \end{aligned} \tag{4.8}$$

By adding (4.6), (4.7) and (4.8), we arrive at the inequality $(B)$. □

Now we are able to provide the proof of Theorem 4.4.1.

*Proof of Theorem 4.4.1.* By combining the results in Lemma 4.3.1 and Lemma 4.4.2, for any feasible policy $\pi$, it holds that

$$\begin{aligned} &\frac{\tau}{T} \cdot (\text{rew}(\pi) + u(\text{con}(\pi))) + \left(1 - \frac{\tau}{T}\right) \cdot B_{\mathcal{V}} \zeta - \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{\tau} r_t^T w_t + u(v^\tau_{\text{avg}})\right] \\ \leq &\frac{1}{T} \cdot (\mathcal{R}_g(T) + \mathcal{R}_\lambda(T) + \zeta \cdot \mathcal{R}_\theta(T)). \end{aligned}$$

Also, it holds that $\mathcal{R}_g(T) \leq \kappa_{\text{risk}} \cdot \mathcal{O}(T^{1-\alpha})$, $\mathcal{R}_\lambda(T) \leq D_v\sqrt{D_\Lambda} \cdot \mathcal{O}(T^{1/2})$, and $\mathcal{R}_\theta(T) \leq D_v\sqrt{D_\Theta} \cdot \mathcal{O}(T^{1/2})$. Therefore, when $\zeta \geq \frac{\text{OPT}}{B_\mathcal{V}}$, it holds that

$$\text{OPT} - \mathbb{E}[r^\tau_{\text{avg}} + u(v^\tau_{\text{avg}})] \leq \kappa_{\text{MD}} \cdot \mathcal{O}(T^{-1/2}) + \kappa_{\text{risk}} \cdot \mathcal{O}(T^{-\alpha}).$$

□

### 4.4.2 Soft Constraints

In this case, we treat the budget consumption feasibility constraint set $\mathcal{V}$ as a soft constraint, that is, we want to minimize the infeasibility of the consumption instead of terminating the online algorithm whenever we violate the constraint. This case allows for the possibility that the starting point of budget consumption is infeasible, for example, when the feasible region $\mathcal{V}$ consists of both lower and upper bound constraints. The following assumption is required for the regret analysis in this case.

**Assumption 4.4.2.** *Let* $\text{OPT}^\epsilon$ *denote the optimal objective value of the following relaxed problem:*

$$\text{OPT}^\epsilon := \sup_\pi \{\text{rew}(\pi) + u(\text{con}(\pi))\}, \quad \text{s.t.}\, d_\mathcal{V}(\text{con}(\pi)) \leq \epsilon.$$

*We assume there exists a constant* $\zeta_{\text{OPT}}$ *such that* $\text{OPT}^\epsilon \leq \text{OPT} + \zeta_{\text{OPT}} \cdot \epsilon$ *for all* $\epsilon > 0$.

The constant $\zeta_{\text{OPT}}$ in Assumption 4.4.2 can be interpreted as a subgradient of the concave function $\text{OPT}^\epsilon$, which can be demonstrated to exist under standard regularity conditions. For example, in the following lemma, we demonstrate that Assumption 4.4.2 holds when $\mathcal{V}$ has a non-empty interior.

**Lemma 4.4.3.** *Suppose that* OPT *is finite and that there exists a policy* $\pi^\circ$ *such that* $\text{con}(\pi^\circ) \in \text{int}(\mathcal{V})$. *Then, there exists a constant* $\zeta'$ *such that*

$$\text{OPT} = \sup_\pi \{\text{rew}(\pi) + u(\text{con}(\pi)) - \zeta' \cdot d_\mathcal{V}(\text{con}(\pi))\}.$$

*Proof.* Let $u^\circ \leftarrow \text{con}(\pi^\circ)$. Since $u^\circ \in \text{int}(\mathcal{V})$, there exists a constant $\epsilon > 0$ such that for all $u'$ satisfying $\|u' - u^\circ\| \leq \epsilon$, it holds that $u' \in \mathcal{V}$. Let $\partial\mathcal{V}$ denote the boundary of the set $\mathcal{V}$, and define $\mathcal{V}^{-\epsilon} \leftarrow \{v \in \mathcal{V} : d_{\partial\mathcal{V}}(v) \geq \epsilon\}$. Consider

$$\text{OPT}^{-\epsilon} = \sup_{\pi:\text{con}(\pi)\in\mathcal{V}^{-\epsilon}} \{\text{rew}(\pi) + u(\text{con}(\pi))\}.$$

Since $\text{con}(\pi^\circ) \in \mathcal{V}^{-\epsilon}$, we know that $\text{OPT}^{-\epsilon}$ is real-valued. Pick a policy $\pi^\dagger$ such that $\text{con}(\pi^\dagger) \in \mathcal{V}^{-\epsilon}$ and $\text{rew}(\pi^\dagger) + u(\text{con}(\pi^\dagger)) \geq \text{OPT}^{-\epsilon} - \epsilon$. Let $v^\dagger \leftarrow \text{con}(\pi^\dagger)$. Now for any $\pi \notin \mathcal{V}$, let $v \leftarrow \text{con}(\pi)$. Pick $\tilde{v} \in \partial\mathcal{V}$ such that $\|v - \tilde{v}\| = d_\mathcal{V}(v)$ and let $\kappa = d_\mathcal{V}(v)/d_{\partial\mathcal{V}}(v^\dagger) \leq d_\mathcal{V}(v)/\epsilon$. Let $\tilde{v}^\dagger \leftarrow v^\dagger + (v - \tilde{v})/\kappa$, since $\|\tilde{v}^\dagger - v^\dagger\| = d_{\partial\mathcal{V}}(v^\dagger)$, it holds that $\tilde{v}^\dagger \in \mathcal{V}$. Also, let $v' \leftarrow \frac{1}{\kappa+1} \cdot (\kappa \cdot \tilde{u}^\dagger + \tilde{v}) \in \mathcal{V}$, and it holds that $v' = \frac{1}{\kappa+1} \cdot (\kappa \cdot u^\dagger + u)$.

Now define a new policy $\pi'$ by $\pi'(x) := \frac{1}{\kappa+1} \cdot (\kappa \cdot \pi^\dagger(x) + \pi(x))$ for any $x \in \mathcal{X}$. The policy $\pi'$ is well-defined since $\mathcal{S}$ is convex and therefore $\pi'(x) \in \mathcal{S}$ for any $x \in \mathcal{X}$. Also, the policy $\pi'$ is feasible since $\mathrm{con}(\pi') = v' \in \mathcal{V}$, and therefore, it holds that $\mathrm{OPT} \geq \mathrm{rew}(\pi') + u(\mathrm{con})(\pi')$. On the other hand, since $u(\cdot)$ is concave, it holds that

$$\mathrm{rew}(\pi') + u(\mathrm{con}(\pi')) \geq \frac{1}{\kappa+1} \cdot (\kappa \cdot [\mathrm{rew}(\pi^\dagger) + u(\mathrm{con}(\pi^\dagger))] + [\mathrm{rew}(\pi) + u(\mathrm{con}(\pi))]).$$

Therefore, it holds that

$$\begin{aligned}
\mathrm{rew}(\pi) + u(\mathrm{con}(\pi)) - \mathrm{OPT} &\leq \kappa \cdot (\mathrm{OPT} - (\mathrm{rew}(\pi^\dagger) + u(\mathrm{con}(\pi^\dagger)))) \\
&= \frac{d_{\mathcal{V}}(\mathrm{con}(\pi))}{d_{\partial\mathcal{V}}(\mathrm{con}(\pi^\circ))} \cdot (\mathrm{OPT} - (\mathrm{OPT}^{-\epsilon} - \epsilon)) \\
&\leq d_{\mathcal{V}}(\mathrm{con}(\pi)) \cdot \left( \frac{\mathrm{OPT} - \mathrm{OPT}^{-\epsilon}}{\epsilon} + 1 \right).
\end{aligned}$$

By setting $\zeta' \leftarrow 1 + (\mathrm{OPT} - \mathrm{OPT}^{-\epsilon})/\epsilon$, for any $\pi \notin \mathcal{V}$, it holds that

$$\mathrm{OPT} \geq \mathrm{rew}(\pi) + u(\mathrm{con}(\pi)) - \zeta' \cdot d_{\mathcal{V}}(\mathrm{con}(\pi)),$$

and we conclude the proof. □

Now given the results in Lemma 4.4.3, for any $\epsilon > 0$, it holds that

$$\begin{aligned}
\mathrm{OPT}^\epsilon &= \sup_{\pi : d_{\mathcal{V}}(\mathrm{con}(\pi)) \leq \epsilon} \{\mathrm{rew}(\pi) + u(\mathrm{con}(\pi))\} \\
&\leq \sup_{\pi} \{\mathrm{rew}(\pi) + u(\mathrm{con}(\pi)) - \zeta' \cdot [d_{\mathcal{V}}(\mathrm{con}(\pi)) - \epsilon])\} \\
&= \sup_{\pi} \{\mathrm{rew}(\pi) + u(\mathrm{con}(\pi)) - \zeta' \cdot d_{\mathcal{V}}(\mathrm{con}(\pi)))\} + \zeta' \cdot \epsilon \\
&\leq \mathrm{OPT} + \zeta' \cdot \epsilon,
\end{aligned}$$

and therefore we show the existence of $\zeta_{\mathrm{OPT}}$.

We are now ready to present the main theorem in the soft constraint case, which demonstrates the convergence rate in terms of both the objective and the distance to feasibility in resource consumption.

**Theorem 4.4.2.** *Suppose that Assumptions 4.3.1, 4.4.1, and 4.4.2 hold. Then, Algorithm 4.2 has the following guarantee:*

$$\mathrm{OPT} - \mathbb{E}[r_{\mathrm{avg}} + u(v_{\mathrm{avg}})] \leq \kappa_{\mathrm{MD}} \cdot \mathcal{O}(T^{-1/2}) + \kappa_{\mathrm{risk}} \cdot \mathcal{O}(T^{-\alpha}).$$

*If additionally the budget penalty parameter $\zeta$ satisfies $\zeta \geq 2(\zeta_{\mathrm{OPT}} + \frac{\sqrt{D_\Lambda}}{\sqrt{D_\Theta}} + \frac{\kappa_{\mathrm{risk}}}{D_v \sqrt{D_\Theta}})$, it holds that*

$$\mathbb{E}[d_{\mathcal{V}}(v_{\mathrm{avg}})] \leq D_v \sqrt{D_\Theta} \cdot \mathcal{O}(T^{-1/2}) + \mathcal{O}(T^{-\alpha}).$$

The following lemma presents the regret from the suboptimality of the dual variables in the soft constraint case.

**Lemma 4.4.4.** *For any feasible policy $\pi(\cdot) : \mathcal{X} \to \mathcal{S}$ and $T \geq 1$, Algorithm 4.2 satisfies*

$$(A): \quad \mathbb{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{T}(V_t\lambda_t)^T(\pi(x_t)-w_t)\right] \leq (-u)(\mathrm{con}(\pi)) - \mathbb{E}[(-u)(v_{\mathrm{avg}})] + \frac{\overline{\mathcal{R}}_\lambda(T)}{T},$$
$$(B): \quad \mathbb{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{T}(V_t\theta_t)^T(\pi(x_t)-w_t)\right] \leq -\mathbb{E}[d_{\mathcal{V}}(v_{\mathrm{avg}})] + \frac{\overline{\mathcal{R}}_\theta(T)}{T}.$$

*Proof.* First, there exists $\lambda \in \Lambda$ such that

$$(-u)(v_{\mathrm{avg}}) = v_{\mathrm{avg}}^T\lambda - (-u)^*(\lambda) = \frac{1}{T}\sum_{t=1}^{T}\left(w_t^TV_t\lambda - (-u)^*(\lambda)\right). \tag{4.9}$$

Next, the Fenchel-Young inequality yields

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\pi(x_t)^TV_t\lambda_t - (-u)^*(\lambda_t)|\mathcal{F}_{t-1}\right] &= \frac{1}{T}\sum_{t=1}^{T}\left(\mathrm{con}(\pi)^T\lambda_t - (-u)^*(\lambda_t)\right)\\ &\leq \frac{1}{T}\sum_{t=1}^{T}(-u)(\mathrm{con}(\pi)) = (-u)(\mathrm{con}(\pi)).\end{aligned} \tag{4.10}$$

Lemma 4.3.2 guarantees that

$$\sum_{t=1}^{T}(\xi_t(\lambda_t) - \xi_t(\lambda)) \leq \overline{\mathcal{R}}_\lambda(T).$$

Given the definition of $\xi(\cdot)$, which is $\xi_t(\lambda') = -v_t^T\lambda' + (-u)^*(\lambda')$, and $v_t = V_t^Tw_t$, it holds that

$$\sum_{t=1}^{T}\left(-w_t^TV_t\lambda_t + (-u)^*(\lambda_t) + w_t^TV_t\lambda - (-u)^*(\lambda)\right) \leq \overline{\mathcal{R}}_\lambda(T). \tag{4.11}$$

Therefore, by adding (4.9), (4.10), and (4.11), it holds that

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{T}(V_t\lambda_t)^T(\pi(x_t)-w_t)\right]\\ \leq&\,\mathbb{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{T}\left[(V_t\lambda_t)^T\pi(x_t) - w_t^TV_t\lambda + (-u)^*(\lambda) - (-u)^*(\lambda_t))\right]\right] + \frac{\overline{\mathcal{R}}_\lambda(T)}{T}\\ \leq&\,(-u)(\mathrm{con}(\pi)) - \mathbb{E}[(-u)(v_{\mathrm{avg}})] + \frac{\overline{\mathcal{R}}_\lambda(T)}{T}.\end{aligned}$$

Applying the same reasoning to the other set of dual variables and using that $\text{con}(\pi) \in \mathcal{V}$, we have

$$\mathbb{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{T}(V_t\theta_t)^T(\pi(x_t)-w_t)\right] \leq d_{\mathcal{V}}(\text{con}(\pi))-\mathbb{E}[d_{\mathcal{V}}(v_{\text{avg}})]+\frac{\overline{\mathcal{R}}_\theta(T)}{T} = -\mathbb{E}[d_{\mathcal{V}}(v_{\text{avg}})]+\frac{\overline{\mathcal{R}}_\theta(T)}{T}.$$

□

Now we are able to provide the proof of Theorem 4.4.1.

*Proof of Theorem 4.4.2.* By combining the results in Lemma 4.3.1 and Lemma 4.4.4, for any feasible policy $\pi$, it holds that

$$\begin{aligned}&\text{rew}(\pi)+u(\text{con}(\pi))-\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}r_t^Tw_t+u(v_{\text{avg}})\right]\\ \leq&\frac{1}{T}\cdot(\mathcal{R}_g(T)+\mathcal{R}_\lambda(T)+\zeta\cdot\mathcal{R}_\theta(T)-\zeta\cdot\mathbb{E}[d_{\mathcal{V}}(v_{\text{avg}})].\end{aligned}$$

Since $\mathbb{E}[d_{\mathcal{V}}(v_{\text{avg}})] \geq 0$, it holds that

$$\text{OPT}-\mathbb{E}[r_{\text{avg}}+u(v_{\text{avg}})] \leq \kappa_{\text{MD}}\cdot\mathcal{O}(T^{-1/2})+\kappa_{\text{risk}}\cdot\mathcal{O}(T^{-\alpha}).$$

Also, when Assumption 4.4.2 holds, it holds that

$$\mathbb{E}[r_{\text{avg}}+u(v_{\text{avg}})]-\text{OPT} \leq \zeta_{\text{OPT}}\cdot\mathbb{E}[d_{\mathcal{V}}(v_{\text{avg}})],$$

and therefore, it holds that

$$(\zeta-\zeta_{\text{OPT}})\cdot\mathbb{E}[d_{\mathcal{V}}(v_{\text{avg}})] \leq \frac{1}{T}\cdot(\mathcal{R}_g(T)+\mathcal{R}_\lambda(T)+\zeta\cdot\mathcal{R}_\theta(T)).$$

If additionally $\zeta$ satisfies $\zeta \geq 2(\zeta_{\text{OPT}}+\frac{\sqrt{D_\Lambda}}{\sqrt{D_\Theta}}+\frac{\kappa_{\text{risk}}}{D_v\sqrt{D_\Theta}})$, it holds that

$$\mathbb{E}[d_{\mathcal{V}}(v_{\text{avg}})] \leq D_v\sqrt{D_\Theta}\cdot\mathcal{O}(T^{-1/2})+\mathcal{O}(T^{-\alpha}).$$

□

To give some intuition of the proofs of Theorem 4.4.1 and Theorem 4.4.2, we remark that the total regret of the online algorithm can be divided into two parts: (i) the regret from the learning of the prediction model, and (ii) the regret from the suboptimality of the dual variables used in each iteration. In the supplementary materials, we present two lemmas to bound each type of regret. Lemma 4.3.1 bounds the regret due to learning, in particular the expected accumulative errors of the online decision $w_t$ due to imperfect predictions, which can be bounded in a sublinear fashion based on Assumption 4.3.1. To bound the regret due to suboptimality of the dual variables, we use the regret bound of online mirror descent method in Lemma 4.3.2 and properties of the Fenchel conjugate functions, which, with a few additional steps, yield Lemma 4.4.2 and Lemma 4.4.4. In the hard and soft cases respectively, these two Lemmas provide guarantees of the decisions $w_t$ from Algorithm 4.2 against any feasible static policy.

### 4.4.3 Trade-off between Regret and Computation Cost

In each iteration of Algorithm 4.2, one essential step is to update the prediction model based on all the previous observations and the current dual variables. Although it may be possible to perform this update efficiently – for example, one could use a warm-starting procedure depending on the structure of the hypothesis class – the decision-maker may still not want to update the prediction model at each iteration, especially if decisions need to be made quickly. To address this issue, we develop a more computationally efficient version of our algorithm, which only updates the prediction model at a sublinear rate, and is formally described as follows.

**Definition 4.4.1.** *For any constant $\beta \geq 1$, the $\beta$-efficient version of Algorithm 4.2 is an algorithm which is same as Algorithm 4.2 but only updates the dual variables and prediction model at iteration $t = \lfloor k^\beta \rfloor$ for all positive integer $k$.*

From the prediction model update frequency of a $\beta$-efficient version of Algorithm 4.2, we notice that a total number of $T^{1/\beta}$ prediction model updates is required. We provide the regret analysis of a $\beta$-efficient version of Algorithm 4.2 in Theorem 4.4.3.

**Theorem 4.4.3.** *In the hard constraints case, suppose that the assumptions of Theorem 4.4.1 hold and consider the $\beta$-efficient version of Algorithm 4.2 for some constant $\beta \in (0, 1]$. Then we have the following guarantee:*

$$\mathrm{OPT} - \mathbb{E}[r^\tau_{\mathrm{avg}} + u(v^\tau_{\mathrm{avg}})] \leq \kappa_{\mathrm{MD}} \cdot \mathcal{O}(T^{-1/2\beta}) + \kappa_{\mathrm{risk}} \cdot \mathcal{O}(T^{-\alpha}).$$

*In the soft constraints case, suppose that the assumptions of Theorem 4.4.2 hold and consider the $\beta$-efficient version of Algorithm 4.2 for some constant $\beta \in (0, 1]$. Then we have the following guarantees:*

$$\mathrm{OPT} - \mathbb{E}[r_{\mathrm{avg}} + u(v_{\mathrm{avg}})] \leq \kappa_{\mathrm{MD}} \cdot \mathcal{O}(T^{-1/2\beta}) + \kappa_{\mathrm{risk}} \cdot \mathcal{O}(T^{-\alpha}),$$

*and if additionally the budget penalty parameter $\zeta$ satisfies $\zeta \geq 2(\zeta_{\mathrm{OPT}} + \frac{\sqrt{D_\Lambda}}{\sqrt{D_\Theta}} + \frac{\kappa_{\mathrm{risk}}}{D_v\sqrt{D_\Theta}})$, it holds that*

$$\mathbb{E}[d_{\mathcal{V}}(v_{\mathrm{avg}})] \leq D_v\sqrt{D_\Theta} \cdot \mathcal{O}(T^{-1/2\beta}) + \mathcal{O}(T^{-\alpha}).$$

*Proof.* When the updating sequence is $t = t_1, \ldots, t_K$, the regret from online mirror descent can be bounded by

$$\sum_{t=1}^{T}(\xi_t(\lambda) - \xi_t(\lambda_t)) \leq \frac{D_\Lambda}{2\eta_\lambda} + \sum_{k=1}^{K}\frac{\eta_\lambda}{2}G^2_\lambda(t_k - t_{k-1})^2.$$

Also, it holds that

$$\sum_{k=1}^{K}(t_k - t_{k-1})^2 = \sum_{k=1}^{K}(\beta k^{\beta-1})^2 = \frac{\beta^2}{2\beta-1} \cdot K^{2\beta-1} = \frac{\beta^2}{2\beta-1} \cdot T^{2-1/\beta}.$$

and therefore by setting $\eta_\lambda \leftarrow \frac{\sqrt{D_\Lambda}}{G_\Lambda T^{1-1/2\beta}}$, we have

$$\mathcal{R}_\lambda(T) = \sum_{t=1}^{T}(\xi_t(\lambda_t) - \xi_t(\lambda)) \leq G_\Lambda\sqrt{D_\Lambda} \cdot \mathcal{O}(T^{1-1/2\beta}).$$

For the same reason we also have

$$\mathcal{R}_\theta(T) = \sum_{t=1}^{T}(\phi_t(\theta_t) - \phi_t(\theta)) \leq G_\Theta\sqrt{D_\Theta} \cdot \mathcal{O}(T^{1-1/2\beta}).$$

Now using the proof in Theorem 4.4.1 and Theorem 4.4.2 again we can get the results in Theorem 4.4.3. □

From the regret analysis, we see that the idea of $\beta$-efficient version of our algorithm is beneficial when the learning of the prediction model has a slower rate, *i.e.*, when $\alpha < \frac{1}{2}$. In this case, we can set $\beta \leftarrow 1/2\alpha$, and the $\beta$-version of Algorithm 4.2 will have a same regret order as the original algorithm, while maintaining a sublinear total number of prediction model updates.

## 4.5 Computational Experiments

We present computational results of synthetic dataset experiments wherein we empirically examine the performance of Algorithm 4.2 using different surrogate loss functions for training prediction models. We focus on two classes of prediction models to represent different levels of model complexity: *(i)* linear models, and *(ii)* two-layer neural networks with 128 neurons in the hidden layer. We compare the performance of the empirical minimizer of the following three different loss functions: *(i)* the previously defined SPO+ loss function, *(ii)* the least squares (squared $\ell_2$) loss function of the linear objective $\|(\hat{r}-\hat{V}\lambda-\zeta\cdot\hat{V}\theta)-(r-V\lambda-\zeta\cdot V\theta)\|_2^2$, and *(iii)* the least squares loss function of predictions $\|\hat{r} - r\|_2^2 + \|\hat{V} - V\|_F^2$. Note that the three loss functions utilize different levels of information: the loss function *(iii)* does not use the dual variables and can be viewed as purely learning the relationship between reward, consumption, and feature vectors. The loss function *(ii)* does not utilize the structure of the decision feasible region $\mathcal{S}$ and can viewed as purely learning the relationship between the linear objectives and feature vectors. We also compare with the following three methods as benchmarks: *(i)* the sample average approximation (SAA) method, where we use the empirical averages of past observations of $r_t, V_t$ as the prediction $\hat{r}_t, \hat{V}_t$ in Algorithm 4.2, *(ii)* the true model, where we use the true (but unknown in practice) conditional expectations $\mathbb{E}[r_t|x_t], \mathbb{E}[V_t|x_t]$ as the prediction, and *(iii)* the hindsight model, where we use the realization $r_t, V_t$ as the prediction. Note that *(ii)* and *(iii)* are not implementable in practice, because *(ii)* uses the unknown true conditional expectations and *(iii)* uses the realized values $r_t, V_t$ that are not available at decision-making time. We expect *(iii)* to perform best and thus we define the "relative regret" of an online algorithm as $1 - \text{OBJ}/\text{OBJ}^*$ where $\text{OBJ} := r_{\text{avg}} + u(v_{\text{avg}})$

is the observed value of the objective function of the online algorithm and OBJ$^*$ is the corresponding value for the hindsight policy.

For all loss functions, we use the Adam method of [47] to train the weight matrices and bias coefficients in the prediction models, and we update the dual variables and prediction models every 10 iterations. For each instance, *e.g.* value of the total time horizon and the polynomial degree, we run 40 independent trials on one core of Intel Xeon Skylake 6230 @ 2.1 GHz.

### 4.5.1 Multi-Dimensional Knapsack Instances

In this section, we consider multi-dimensional knapsack problem instances, where the goal is to maximize total reward collected. There is no utility function, the resource consumption feasible region is $\mathcal{V} = \{v : v \leq b \cdot e\}$ for constant $b > 0$ and the online algorithm must terminates immediately when any of the resource constraints are violated. In our simulations, the relationship between the true reward vector $r$, true resource consumption matrix $V$, and its context vector $x$ is given by $\text{vec}(r, V) \leftarrow \xi^{\text{deg}}(Wx) \odot \epsilon$, where $\text{vec}(\cdot)$ is the matrix vectorization function, $\xi^{\text{deg}}$ is a polynomial kernel mapping of degree deg, $W \in \mathbb{R}^{d(m+1)\times p}$ is a fixed weight matrix, and $\epsilon \in \mathbb{R}^{d(m+1)}$ is a multiplicative noise term.

The detailed data generation process is as followed. In this experiment, we set the dimension of the feature vector $p = 5$, the dimension of decision vector $d = 10$, and the dimension of the resource vector $m = 3$. We first generate the weight matrix $W \in \mathbb{R}^{d(m+1)\times p}$, whereby each entry of $W$ is a Bernoulli random variable with the probability $\mathbb{P}(B_{jk} = 1) = \frac{1}{2}$. We then generate the arrivals $\{(x_i, r_i, V_i)\}_{i=1}^{p}$ independently by the following procedure:

1. Generate the feature vector $x$ from a standard multivariate normal distribution, namely $x \sim \mathcal{N}(0, I_p)$.

2. Generate the vectorization of the reward vector $r$ and the resource consumption matrix $V$ according to
$$\text{vec}(r, V)_j \leftarrow \left[1 + \left(1 + \frac{W_j^T x}{\sqrt{p}}\right)^{\text{deg}}\right] \epsilon_j,$$
for $j = 1, \ldots, d(m+1)$, where $W_j$ is the $j$-th row of matrix $W$. Here deg is the fixed degree parameter and $\epsilon_j$, the multiplicative noise term, is a random variable which independently generated from the uniform distribution $[1 - \bar{\epsilon}, 1 + \bar{\epsilon}]$ for a fixed noise half width $\bar{\epsilon} \geq 0$. In particular, $\bar{\epsilon}$ is set to 0 for "no noise" instances and 0.5 for "moderate noise" instances.

We set the polynomial degree to 6 in this experiment, and we run 40 independent trials for each value of the time horizon length. Figure 4.1 displays the empirical performance of each method. We observe that when the hypothesis class is linear predictors, *i.e.*, the ground truth model is not in the hypothesis class, the pure prediction error method has

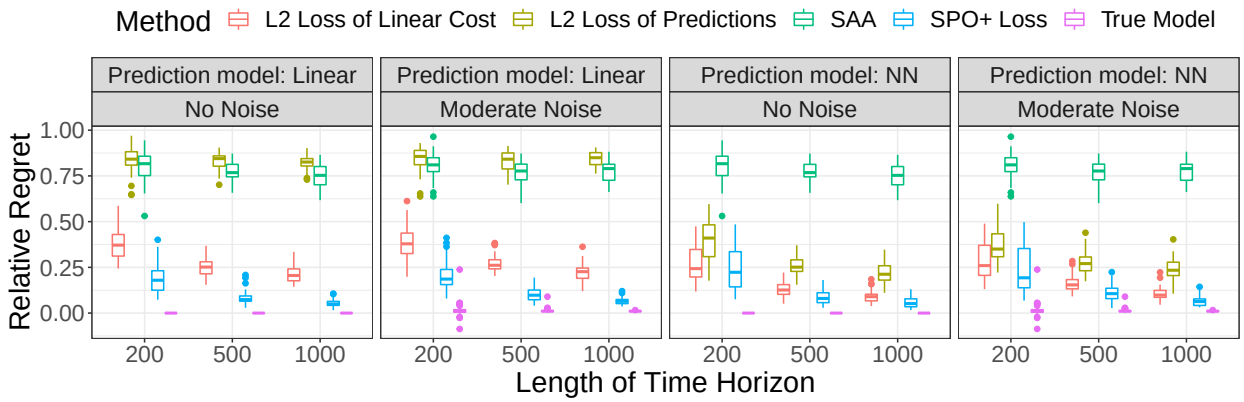



Figure 4.1: Relative regret for different loss functions on multi-dimensional knapsack instances.

similar performance as naive SAA, while least squares applied to the linear cost performs slightly better. When the hypothesis class is a neural net, the pure prediction method does better. In all cases, the SPO+ loss performs best and closest to the true model. These results demonstrate that a loss function that properly accounts for the dual variables can improve performance, but performance is improved twofold by a loss function that accounts for the dual variables *and* the underlying structure of the decision feasible region $\mathcal{S}$.

### 4.5.2 Longest Path Instances

In this section, we consider a longest path problem on a $4 \times 4$ directed grid network with edges pointing north and east, and the goal is to go from the southwest corner to the northeast corner while maximizing the rewards collected along each edge. In this case, the feasible region $\mathcal{S}$ can be modeled as the convex hull of all possible routes. We assume there is no learning in the consumption matrix, *i.e.*, the consumption is just the decision itself, namely $V_t = I_d$, i.e., $v_t = w_t$. Also, we would like to not utilize any edge too frequently and model this idea by setting the resource consumption feasible set as $\mathcal{V} = \{v : v \leq 0.6 \cdot e\}$, which is a soft constraint, and letting the utility function be $u(v) = \sum_i v_i(1 - v_i)$.

The detailed data generation process in the longest path instances is as followed. In this experiment, we set the dimension of the feature vector $p = 5$. Also, since the graph is a $4 \times 4$ grid, the dimension of both decision and resource consumption vector will be $d = m = 24$, which is the number of edges in the graph. Since the resource consumption matrix is always the identity matrix, we only need to generate the reward vector based on the feature. Therefore, the weight matrix is $W \in \mathbb{R}^{d \times p}$. The remaining part is the same as the data generation process in the multi-dimensional knapsack instances. We set the total number of arrivals to 1000 in this experiment. Figure 4.2 displays the empirical performance

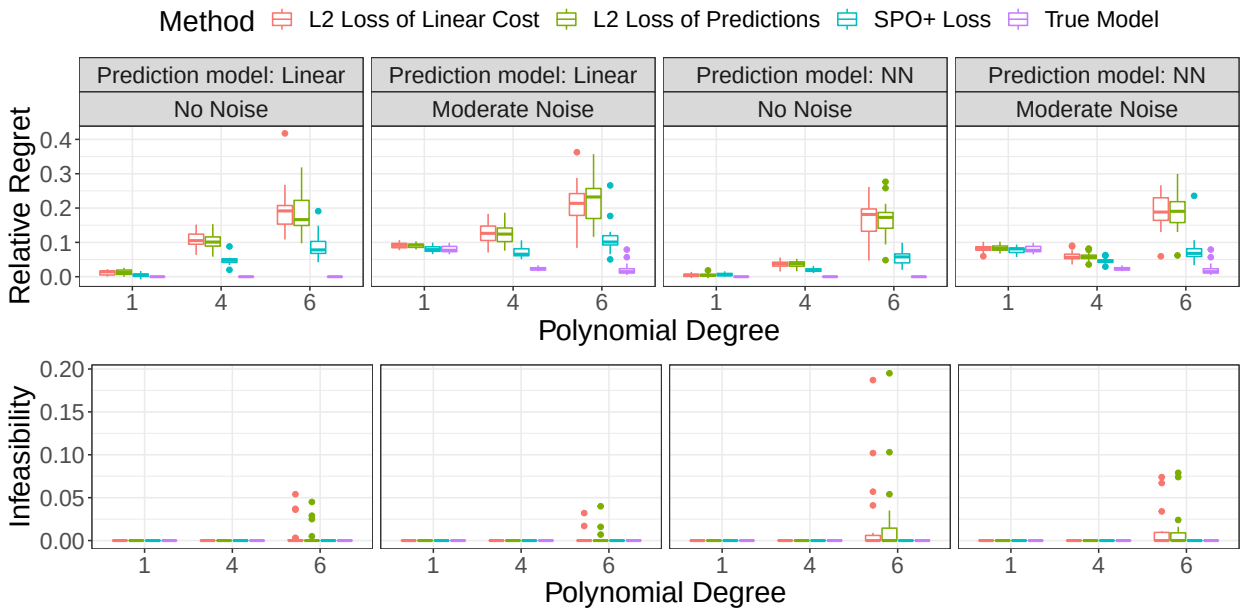



Figure 4.2: Relative regret and infeasibility for different loss functions on shortest path instances.

of each method. We observe that the SPO+ loss which accounts for both dual variables and the decision feasible region $\mathcal{S}$ dominates all cases, and it is more beneficial when the polynomial degree is higher.

# Bibliography


[1] Boushra Abbas, Hédy Attouch, and Benar F Svaiter. "Newton-like dynamics and forward-backward methods for structured monotone inclusions in Hilbert spaces". In: *Journal of Optimization Theory and Applications* 161.2 (2014), pp. 331–360.

[2] Shipra Agrawal and Nikhil Devanur. "Linear contextual bandits with knapsacks". In: *Advances in Neural Information Processing Systems* 29 (2016).

[3] Shipra Agrawal and Nikhil R Devanur. "Bandits with concave rewards and convex knapsacks". In: *Proceedings of the fifteenth ACM conference on Economics and computation*. 2014, pp. 989–1006.

[4] Shipra Agrawal and Nikhil R Devanur. "Fast algorithms for online stochastic convex programming". In: *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2014, pp. 1405–1424.

[5] Shipra Agrawal and Navin Goyal. "Thompson sampling for contextual bandits with linear payoffs". In: *International conference on machine learning*. PMLR. 2013, pp. 127–135.

[6] Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. "A dynamic near-optimal algorithm for online linear programming". In: *Operations Research* 62.4 (2014), pp. 876–890.

[7] Felipe Alvarez and Hedy Attouch. "An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping". In: *Set-Valued Analysis* 9.1-2 (2001), pp. 3–11.

[8] Felipe Alvarez et al. "A second-order gradient-like dissipative dynamical system with Hessian-driven damping.: Application to optimization and mechanics". In: *Journal de mathématiques pures et appliquées* 81.8 (2002), pp. 747–779.

[9] Hedy Attouch, Maicon Marques Alves, and Benar Fux Svaiter. "A Dynamic Approach to a Proximal-Newton Method for Monotone Inclusions in Hilbert Spaces, with Complexity $O(1/n^2)$". In: *Journal of Convex Analysis* 23.1 (2016), pp. 139–180.

[10] Hedy Attouch and Benar Fux Svaiter. "A continuous dynamical Newton-like approach to solving monotone inclusions". In: *SIAM Journal on Control and Optimization* 49.2 (2011), pp. 574–598.

[11] Hedy Attouch et al. "Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity". In: *Mathematical Programming* 168.1-2 (2018), pp. 123–175.

[12] Francis R Bach, David Heckerman, and Eric Horvitz. "Considering cost asymmetry in learning classifiers". In: *Journal of Machine Learning Research* 7.Aug (2006), pp. 1713–1741.

[13] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. "Bandits with knapsacks". In: *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE. 2013, pp. 207–216.

[14] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. "Bandits with knapsacks". In: *Journal of the ACM (JACM)* 65.3 (2018), pp. 1–55.

[15] Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. "Resourceful contextual bandits". In: *Conference on Learning Theory*. PMLR. 2014, pp. 1109–1134.

[16] Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. "The best of many worlds: Dual mirror descent for online allocation problems". In: *arXiv preprint arXiv:2011.10124* (2020).

[17] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. "Convexity, classification, and risk bounds". In: *Journal of the American Statistical Association* 101.473 (2006), pp. 138–156.

[18] Peter L Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482.

[19] Anil K Bera and Sung Y Park. "Optimal portfolio diversification using the maximum entropy principle". In: *Econometric Reviews* 27.4-6 (2008), pp. 484–512.

[20] Dimitris Bertsimas and Nathan Kallus. "From predictive to prescriptive analytics". In: *Management Science* 66.3 (2020), pp. 1025–1044.

[21] Radu Ioan Bot and Ernö Robert Csetnek. "Second order forward-backward dynamical systems for monotone inclusion problems". In: *SIAM Journal on Control and Optimization* 54.3 (2016), pp. 1423–1443.

[22] J. C. Butcher. *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods*. New York, NY, USA: Wiley-Interscience, 1987. ISBN: 0-471-91046-5.

[23] Andrew R Conn, Nicholas IM Gould, and Ph L Toint. *Trust region methods*. Vol. 1. Siam, 2000.

[24] Nikhil R Devanur et al. "Near optimal online algorithms and fast approximation algorithms for resource allocation problems". In: *Proceedings of the 12th ACM conference on Electronic commerce*. 2011, pp. 29–38.

[25] Priya Donti, Brandon Amos, and J. Zico Kolter. "Task-based End-to-end Model Learning in Stochastic Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[26] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.

[27] Bradley Efron et al. "Least angle regression". In: *The Annals of statistics* 32.2 (2004), pp. 407–499.

[28] Othman El Balghiti et al. "Generalization Bounds in the Predict-then-Optimize Framework". In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[29] Adam N Elmachtoub and Paul Grigas. "Smart "predict, then optimize"". In: *Management Science* (2021).

[30] Alexander Estes and Jean-Philippe Richard. "Objective-Aligned Regression for Two-Stage Linear Programs". In: *Available at SSRN 3469897* (2019).

[31] Kris Johnson Ferreira, David Simchi-Levi, and He Wang. "Online network revenue management using thompson sampling". In: *Operations research* 66.6 (2018), pp. 1586–1602.

[32] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1 (2010), p. 1.

[33] Carlos E Garcia, David M Prett, and Manfred Morari. "Model predictive control: Theory and practice—A survey". In: *Automatica* 25.3 (1989), pp. 335–348.

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

[35] Joachim Giesen, Martin Jaggi, and Sören Laue. "Approximating parameterized convex optimization problems". In: *ACM Transactions on Algorithms (TALG)* 9.1 (2012), p. 10.

[36] Joachim Giesen, Martin Jaggi, and Sören Laue. "Regularization paths with guarantees for convex semidefinite optimization". In: *Artificial Intelligence and Statistics*. 2012, pp. 432–439.

[37] Todd R Golub et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". In: *science* 286.5439 (1999), pp. 531–537.

[38] Jürgen Guddat, F Guerra Vazquez, and Hubertus Th Jongen. *Parametric optimization: singularities, pathfollowing and jumps*. Springer, 1990.

[39] Trevor Hastie et al. "The entire regularization path for the support vector machine". In: *Journal of Machine Learning Research* 5.Oct (2004), pp. 1391–1415.

[40] Chin Pang Ho and Grani A Hanasusanto. "On data-driven prescriptive analytics with side information: A regularized nadaraya-watson approach". In: *URL: http://www.optimization-online. org/DB FILE/2019/01/7043. pdf* (2019).

[41] Yichun Hu, Nathan Kallus, and Xiaojie Mao. "Fast Rates for Contextual Linear Optimization". In: *arXiv preprint arXiv:2011.03030* (2020).

[42] Rodolphe Jenatton, Jim Huang, and Cédric Archambeau. "Adaptive algorithms for online convex optimization with long-term constraints". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 402–411.

[43] Michel Journée et al. "Generalized power method for sparse principal component analysis". In: *Journal of Machine Learning Research* 11.Feb (2010), pp. 517–553.

[44] Yi-hao Kao, Benjamin Roy, and Xiang Yan. "Directed regression". In: *Advances in Neural Information Processing Systems* 22 (2009), pp. 889–897.

[45] Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi. "Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients". In: *arXiv preprint arXiv:1806.00413* (2018).

[46] Seung-Jean Kim et al. "An interior-point method for large-scale $\ell_1$-regularized least squares". In: *IEEE journal of selected topics in signal processing* 1.4 (2007), pp. 606–617.

[47] Diederik P Kingma and Jimmy Lei Ba. "Adam: A method for stochastic gradient descent". In: *ICLR: International Conference on Learning Representations*. 2015, pp. 1–15.

[48] James Kotary et al. "End-to-End Constrained Optimization Learning: A Survey". In: *arXiv preprint arXiv:2103.16378* (2021).

[49] Xiaocheng Li and Yinyu Ye. "Online linear programming: Dual convergence, new algorithms, and regret bounds". In: *Operations Research* (2021).

[50] Nikolaos Liakopoulos et al. "Cautious regret minimization: Online optimization with long-term budget constraints". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3944–3952.

[51] Heyuan Liu and Paul Grigas. "Risk bounds and calibration for a smart predict-then-optimize method". In: *Advances in Neural Information Processing Systems* 34 (2021).

[52] Alfonso Lobos, Paul Grigas, and Zheng Wen. "Joint Online Learning and Decision-making via Dual Mirror Descent". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 7080–7089.

[53] Gaëlle Loosli, Gilles Gasso, and Stéphane Canu. "Regularization Paths for $\nu$-SVM and $\nu$-SVR". In: *International Symposium on Neural Networks*. Springer. 2007, pp. 486–496.

[54] Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. "Trading regret for efficiency: online convex optimization with long term constraints". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 2503–2528.

[55] Julien Mairal and Bin Yu. "Complexity analysis of the lasso regularization path". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress. 2012, pp. 1835–1842.

[56] Harry Markowitz. "Portfolio Selection". In: *The Journal of Finance* 7.1 (1952), pp. 77–91. ISSN: 00221082, 15406261. URL: http://www.jstor.org/stable/2975974.

[57] Donald W Marquardt. "An algorithm for least-squares estimation of nonlinear parameters". In: *Journal of the society for Industrial and Applied Mathematics* 11.2 (1963), pp. 431–441.

[58] Pascal Massart, Élodie Nédélec, et al. "Risk bounds for statistical learning". In: *The Annals of Statistics* 34.5 (2006), pp. 2326–2366.

[59] Andreas Maurer. "A vector-contraction inequality for rademacher complexities". In: *International Conference on Algorithmic Learning Theory*. Springer. 2016, pp. 3–17.

[60] Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. "Sparsenet: Coordinate descent with nonconvex penalties". In: *Journal of the American Statistical Association* 106.495 (2011), pp. 1125–1138.

[61] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. "Spectral regularization algorithms for learning large incomplete matrices". In: *Journal of machine learning research* 11.Aug (2010), pp. 2287–2322.

[62] Jorge J Moré. "The Levenberg-Marquardt algorithm: implementation and theory". In: *Numerical analysis*. Springer, 1978, pp. 105–116.

[63] Eugene Ndiaye et al. "Safe Grid Search with Optimal Complexity". In: *International Conference on Machine Learning*. 2019, pp. 4771–4780.

[64] Yurii Nesterov. *Implementable tensor methods in unconstrained convex optimization*. Tech. rep. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2018.

[65] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. Vol. 13. Siam, 1994.

[66] Yurii Nesterov and Boris T Polyak. "Cubic regularization of Newton method and its global performance". In: *Mathematical Programming* 108.1 (2006), pp. 177–205.

[67] Nam Ho-Nguyen and Fatma Kılınç-Karzan. "Risk guarantees for end-to-end prediction and optimization processes". In: *Management Science* (2022).

[68] Pascal M Notz and Richard Pibernik. "Prescriptive analytics for flexible capacity management". In: *Available at SSRN 3387866* (2019).

[69] Michael R Osborne, Brett Presnell, and Berwin A Turlach. "A new approach to variable selection in least squares problems". In: *IMA journal of numerical analysis* 20.3 (2000), pp. 389–403.

[70] Anton Osokin, Francis Bach, and Simon Lacoste-Julien. "On structured prediction theory with calibrated convex surrogate losses". In: *arXiv preprint arXiv:1703.02403* (2017).

[71] Aldo Pacchiano et al. "Stochastic bandits with linear constraints". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2827–2835.

[72] Roman A Polyak. "Regularized Newton method for unconstrained convex optimization". In: *Mathematical programming* 120.1 (2009), pp. 125–145.

[73] Daniel Ralph. "Global convergence of damped Newton's method for nonsmooth equations via the path search". In: *Mathematics of Operations Research* 19.2 (1994), pp. 352–389.

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

[75] Saharon Rosset. "Following Curved Regularized Optimization Solution Paths". In: *Advances in Neural Information Processing Systems 17*. Ed. by L. K. Saul, Y. Weiss, and L. Bottou. MIT Press, 2005, pp. 1153–1160. URL: http://papers.nips.cc/paper/2600-following-curved-regularized-optimization-solution-paths.pdf.

[76] Saharon Rosset and Ji Zhu. "Piecewise linear regularized solution paths". In: *The Annals of Statistics* (2007), pp. 1012–1030.

[77] Damien Scieur et al. "Integration methods and optimization algorithms". In: *Advances in Neural Information Processing Systems*. 2017, pp. 1109–1118.

[78] L.R. Scott. *Numerical Analysis*. Princeton University Press, 2011. ISBN: 9781400838967. URL: https://books.google.com/books?id=SfCjL%5C_5AaRQC.

[79] Shai Shalev-Shwartz et al. "Online learning and online convex optimization". In: *Foundations and Trends® in Machine Learning* 4.2 (2012), pp. 107–194.

[80] Ingo Steinwart. "How to compare different loss functions and their risks". In: *Constructive Approximation* 26.2 (2007), pp. 225–287.

[81] Weijie Su, Stephen Boyd, and Emmanuel Candes. "A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2510–2518.

[82] Ambuj Tewari and Peter L Bartlett. "On the Consistency of Multiclass Classification Methods." In: *Journal of Machine Learning Research* 8.5 (2007).

[83] Alberto Vera, Siddhartha Banerjee, and Itai Gurvich. "Online allocation and pricing: Constant regret via bellman inequalities". In: *Operations Research* 69.3 (2021), pp. 821–840.

[84] Zhaoran Wang, Han Liu, and Tong Zhang. "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems". In: *Annals of statistics* 42.6 (2014), p. 2164.

[85] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. "A variational perspective on accelerated methods in optimization". In: *proceedings of the National Academy of Sciences* 113.47 (2016), E7351–E7358.

[86] Ming Yuan and Hui Zou. "Efficient global approximation of generalized nonlinear $\ell_1$-regularized solution paths and its applications". In: *Journal of the American Statistical Association* 104.488 (2009), pp. 1562–1574.

[87] Jingzhao Zhang et al. "Direct Runge-Kutta discretization achieves acceleration". In: *Advances in Neural Information Processing Systems*. 2018, pp. 3900–3909.

[88] Tong Zhang. "Statistical analysis of some multi-category large margin classification methods". In: *Journal of Machine Learning Research* 5.Oct (2004), pp. 1225–1251.

[89] Tong Zhang et al. "Statistical behavior and consistency of classification methods based on convex risk minimization". In: *The Annals of Statistics* 32.1 (2004), pp. 56–85.

[90] Hua Zhou and Kenneth Lange. "Path following in the exact penalty method of convex programming". In: *Computational optimization and applications* 61.3 (2015), pp. 609–634.

[91] Hua Zhou and Yichao Wu. "A generic path algorithm for regularized statistical estimation". In: *Journal of the American Statistical Association* 109.506 (2014), pp. 686–699.

# Appendix A

# Supplement to Chapter 2

## A.1 Additional Proofs

### A.1.1 Proof of Proposition 2.2.1

For fixed $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, we have $\frac{\xi(\lambda)}{\lambda}$ is a constant and its absolute value is no larger than $C$. Let $H_i = \nabla^2 f(x_1) + \lambda \nabla^2 \Omega(x_1)$ and $g_i = \nabla f(x_i)$ for $i = 1, 2$. Since $f(\cdot)$ is $\mu$-strongly convex and $\Omega(\cdot)$ is $\sigma$-strongly convex, it holds that $\|H_i\| \geq \mu + \lambda_{\min}\sigma$ for $i = 1, 2$. Since $f(\cdot)$ is $L$-Lipschitz continuous, we have $\|g_2\| \leq L$. Also, let $v_1 = H_1^{-1} g_1$, $v_2 = H_2^{-1} g_2$, and $v_1' = H_1^{-1} g_2$. t holds that

$$\|v_1 - v_1'\| = \|H_1^{-1}(g_1 - g_2)\| \leq \left\|H_1^{-1}\right\| \cdot \|g_1 - g_2\| \leq \frac{1}{\mu + \lambda_{\min}\sigma} \cdot L\|x_1 - x_2\|.$$

Also, since $\|H_1 - H_2\| \leq \|\nabla^2 f(x_1) - \nabla^2 f(x_2)\| + \lambda\|\nabla^2 \Omega(x_1) - \nabla^2 \Omega(x_2)\| \leq L(1 + \lambda_{\max})\|x_1 - x_2\|$ and $\|H_1 - H_2\| = \|H_1(H_1^{-1} - H_2^{-1})H_2\| \leq \|H_1\| \cdot \|H_1^{-1} - H_2^{-1}\| \cdot \|H_2\|$, it holds that $\left\|H_1^{-1} - H_2^{-1}\right\| \leq \frac{L(1+\lambda_{\max})\|x_1 - x_2\|}{(\mu + \lambda_{\min}\sigma)^2}$. Therefore, we have

$$\|v_1' - v_2\| = \left\|\left(H_1^{-1} - H_2^{-1}\right) g_2\right\| \leq \left\|H_1^{-1} - H_2^{-1}\right\| \cdot \|g_2\| \leq \frac{L^2(1 + \lambda_{\max})\,\|x_1 - x_2\|}{(\mu + \lambda_{\min}\sigma)^2}.$$

To conclude, since $v(x_1, \lambda) - v(x_2, \lambda) = \frac{\xi(\lambda)}{\lambda} \cdot (v_1 - v_2)$, it holds that

$$\|v(x_1, \lambda) - v(x_2, \lambda)\| \leq \left(\frac{LC}{\mu + \lambda_{\min}\sigma} + \frac{L^2 C(1 + \lambda_{\max})}{(\mu + \lambda_{\min}\sigma)^2}\right) \cdot \|x_1 - x_2\|.$$

### A.1.2 Proof of Lemma 2.4.1

In the following residual analysis, we will work with high-order directional derivatives of $f(\cdot)$ and $\Omega(\cdot)$. We now introduce definition of the directional derivatives. 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

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

For detailed properties of directional derivatives we refer readers to [64]. We start with computational guarantees of Taylor approximation on functions with Lipschitz continuous high-order derivatives. The following lemma guarantees the accuracy of Taylor expansion.

**Lemma A.1.1** ((1.5) and (1.6) in [64])**.** *Let function $\phi(\cdot)$ be convex and $p$-times differentiable. Suppose $p$-th order derivative of $\phi(\cdot)$ are $L_p$ Lipschitz continuous. Let $\Phi_{x,p}(\cdot)$ denote the Taylor approximation of function $\phi(\cdot)$ at $x$:*

$$\Phi_{x,p}(y) := f(x) + \sum_{i=1}^{p} \frac{1}{i!} D^i \phi(x)[y - x]^i.$$

*Then we have the following guarantees:*

$$|\phi(y) - \Phi_{x,p}(y)| \leq \frac{L_p}{(p+1)!} \|y - x\|^{p+1}, \quad \|\nabla\phi(y) - \nabla\Phi_{x,p}(y)\| \leq \frac{L_p}{p!} \|y - x\|^p.$$

**Condition A.1.1.** *Suppose step-size $h$ satisfies that $h \leq \min\left\{0.2, \frac{\tilde{\mu}}{8L(1+G)}\right\}$.*

For simplicity, we use $(\tilde{x}, \tilde{\lambda})$ to represent $(x_{\text{next}}, \lambda_{\text{next}})$. We will begin the complexity analysis of trapezoid method by the following two lemmas which provide proper upper bounds the norm of direction $\|d_1\|$ and the difference between $d_1$ and $d_2$.

**Lemma A.1.2.** *Suppose $\sigma \geq 1$, $x \in S_{x_0}$, $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, $h > 0$ 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*

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

*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 $(1 + \lambda) \|d_1\| \leq \frac{1+\lambda}{\lambda} G \leq 2$. Also when $\lambda \leq 1$, it holds that

$$(1 + \lambda) \|d_1\| \leq (1 + \lambda) \left(\left\|H^{-1} \lambda \nabla\Omega(x)\right\| + \left\|H^{-1}(\nabla f(x) + \lambda \nabla\Omega(x))\right\|\right) \leq 2(G + 1).$$

□

**Lemma A.1.3.** *Suppose Assumption 2.4.1 and Condition A.1.1 holds. Let $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^+$, and let $x_i, \lambda_i, d_i, i \in \{1, 2\}$ are generated by trapezoid update scheme defined in (2.13), we have*

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

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

$$\|d_2 - d_1\| \leq 2hL\tau \left(\|d_1\| + \|d_1\|^2\right).$$

*Proof.* Using the definition of $d_1, d_2$ we have

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

where $\|(R)\| = \|\nabla f(x_1) - \nabla f(x_2) - \nabla^2 f(x_1)(x_1 - x_2)\| \leq \frac{h^2}{2} \cdot L \|d_1\|^2$. Also, it holds that $\|\nabla^2 f(x_1)(x_1 - x_2)\| = h \|\nabla^2 f(x_1) d_1\| \leq hL \|d_1\|$, and that

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

Hence, it holds that

$$\begin{aligned}&\tilde{\mu} \|d_2\| - \tilde{\mu} \|d_1\| \leq \tilde{\mu} \|d_2 - d_1\| \leq \|\tilde{H}_1(d_2 - d_1)\| \\ \leq&\frac{h^2}{2} \cdot L \|d_1\|^2 + hL \|d_1\| + h\left(L(\lambda + 1) \|d_1\| + \lambda L\right) \|d_2\|.\end{aligned} \tag{A.3}$$

When $h$ satisfies Condition A.1.1, it holds that $h\left(L(\lambda + 1) \|d_1\| + \lambda L\right) \leq \frac{\tilde{\mu}}{3}$ and $\frac{h^2}{2} \cdot L \|d_1\| + hL \leq \frac{\tilde{\mu}}{3}$. Apply them to (A.3), it holds that $\frac{2}{3}\tilde{\mu} \|d_2\| \leq \frac{4}{3}\tilde{\mu} \|d_1\|$ and it implies that $\|d_2\| \leq 2 \|d_1\|$. Apply it to (A.3), it holds that

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

Hence, it holds that

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

□

Based on the results of Lemma A.1.2 and Lemma A.1.3, the following theorem analyze one-step residual accumulation of the trapezoid update in (2.13).

*Proof of Lemma 2.4.1.* We will begin the local residual analysis by estimating the difference between $\nabla F_{\tilde{\lambda}}(\tilde{x})$ and $\frac{\tilde{\lambda}}{\lambda} \cdot \nabla F_\lambda(x)$. For simplicity, let $x, \tilde{x}, \lambda, \tilde{\lambda}$ denote $x_k, x_{k+1}, \lambda_k, \lambda_{k+1}$ in trapezoid update. After rearrangement, we have

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

We will approach the result in (2.14) by splitting and rearranging terms in $(A)$, $(B)$ and $(C)$. From Lemma A.1.1, it holds that

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

From the update (2.13), it holds that $(A') = \underbrace{h\nabla^2 f(x)d_1}_{(A1)} + \underbrace{\frac{h}{2}\nabla^2 f(x)(d_2-d_1)}_{(A2)}$. For $(B)$, using Lemma A.1.1 and based on update (2.13) we have

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

where $\|(RB)\| = \tilde{\lambda}\|\nabla\Omega(\tilde{x}) - \nabla\Omega(x) - \nabla^2\Omega(x)(\tilde{x}-x) - \frac{1}{2}D^3\Omega(x)[\tilde{x}-x]^2\| \leq \frac{h^3\tilde{\lambda}L}{6}\|\tilde{d}\|^3$. Also, $(C) = \left(h - \frac{h^2}{2}\right)\nabla f(x) = \underbrace{h\nabla f(x)}_{(C1)} - \underbrace{\frac{h^2}{2}\nabla f(x)}_{(C2)}$. By the definition of $d_1$, we have

$$(A1) + (B1) + (C1) = h\nabla^2 f(x)d_1 + h\lambda\nabla^2\Omega(x)d_1 + h\nabla f(x) = 0. \tag{A.4}$$

Using Lemma A.1.3, we have

$$\begin{aligned}(A2) + (B2) = &\underbrace{\frac{h}{2}\nabla^2 f(x)(x_1-x_2)}_{(D1)} + \underbrace{\frac{h}{2}(\nabla^2 f(x_1) - \nabla^2 f(x_2))d_2}_{(D2)}\\ &+ \underbrace{\frac{h}{2}(\lambda_1\nabla^2\Omega(x_1) - \lambda_2\nabla^2\Omega(x_2))d_2}_{(D3)} + (RD),\end{aligned} \tag{A.5}$$

where $\|(RD)\| \leq \frac{h}{2}\cdot\frac{L}{2}\|x_1-x_2\|^2 = \frac{h^3}{4}L\|d_1\|^2$. Furthermore, it holds that

$$(D1) + (C2) = -\frac{h^2}{2}\nabla^2 f(x)d_1 - \frac{h^2}{2}\nabla f(x) = \underbrace{\frac{h^2}{2}\lambda\nabla^2\Omega(x)d_1}_{(E1)}, \tag{A.6}$$

and

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

where

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

We further have $(D3) = \underbrace{\frac{h}{2}\lambda_2\left(\nabla^2\Omega(x_1) - \nabla^2\Omega(x_2)\right) d_2}_{(E2)} + \underbrace{\frac{h}{2}(\lambda_1 - \lambda_2)\nabla^2\Omega(x_1) d_2}_{(E3)}$, and

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

where

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

and

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

Hence, it holds that

$$\begin{aligned}
\|(R)\| \leq & \|(RA)\| + \|(RB)\| + \|(RD)\| + \|(R1)\| + \|(R2)\| \\
& + \|(R3)\| + \|(R4)\| + \|(R5)\| + \|(R6)\| \\
\leq & h^3 \cdot \frac{L}{6} \left\| \frac{d_1 + d_2}{2} \right\|^3 + h^3 \cdot \frac{\tilde{\lambda} L}{6} \left\| \frac{d_1 + d_2}{2} \right\|^3 + h^3 \cdot \frac{L}{6} \|d_1\|^2 \\
& + h^3 \cdot \frac{L}{4} \|d_1\|^2 \|d_2\| + h^2 \cdot \frac{L}{8} \|d_1 - d_2\|^2 + h^4 \cdot \frac{\lambda L}{4} \left\| \frac{d_1 + d_2}{2} \right\|^2 \\
& + h^2 \cdot \frac{\lambda_2 L}{8} \|d_1 - d_2\|^2 + h^3 \cdot \frac{\lambda_2 L}{4} \|d_1\|^2 \|d_2\| + h^3 \cdot \frac{\lambda L}{4} \|d_1 - d_2\| \\
\leq & h^3 \cdot \frac{L}{3} \|d_1\|^3 + h^3 \cdot \frac{L\lambda}{3} \|d_1\|^3 + h^3 \cdot \frac{L}{6} \|d_1\|^2 + h^3 \cdot \frac{L}{2} \|d_1\|^3 \\
& + h^2 \cdot \frac{L}{8} \|d_1 - d_2\|^2 + h^4 \cdot \frac{\lambda L}{2} \|d_1\|^2 + h^2 \cdot \frac{\lambda L}{8} \|d_1 - d_2\|^2 \\
& + h^3 \cdot \frac{\lambda L}{2} \|d_1\|^3 + h^3 \cdot \frac{\lambda L}{4} \|d_1 - d_2\| \\
\leq & h^3 L(1 + \lambda) \left( \|d_1\|^3 + \|d_1\|^2 + \|d_1\| \right) + h^2 \cdot \frac{L(1 + \lambda)}{8} \|d_1 - d_2\|^2 .
\end{aligned} \tag{A.7}$$

Apply the result in Lemma A.1.3 into (A.7), we can further get

$$\|d_1 - d_2\|^2 \leq \left( 2hL\tau \left( \|d_1\| + \|d_1\|^2 \right) \right)^2 \leq 8h^2 L^2 \tau^2 \left( \|d_1\|^2 + \|d_1\|^4 \right). \tag{A.8}$$

Also, by applying (A.1) into (A.7) and (A.8), we have

$$\|(R)\| \leq h^3 \cdot 3L(1 + G)^3 + h^4 \cdot 2L^3 \tau^2 (1 + G)^4.$$

□

## A.2 Feasibility in Moment Matching Problem

**Lemma A.2.1.** *Let $x(\cdot) : [\lambda_{\min}, \lambda_{\max}] \to \mathbb{R}^p$ be the exact solution path of $(P)$, then it holds that $x(\lambda)$ is in the relative interior of $S = \{x \in \mathbb{R}^{p+1} : x \geq 0, \mathbf{1}_{p+1}^T x = 1\}$ for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. Also, when $h \leq \frac{\lambda_{\min}(\mu + \lambda_{\min})}{4LG}$, the sequence $\{y_k\}_{k=0}^K$ generated by Algorithms 2.2 and 2.3 on problem $(P')$ satisfies $y_k \in \operatorname{int}(S')$ for all $k = 0, \ldots, K$, and the corresponding sequence $\{x_k\}_{k=0}^K$ satisfies $x_k$ is in the relative interior of $S$ for all $k = 0, \ldots, K$. Under same condition, it holds that $\hat{x}(\cdot)$ generated by (2.11) is a subset of the relative interior of $S$ for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. Moreover, the condition $h \leq \frac{\lambda_{\min}(\mu + \lambda_{\min})}{4LG}$ is automatically satisfied when $\epsilon \leq \frac{(f(x_0) - f^*)\lambda_{\min}}{4G}$ with the choice of step-size in Theorem 2.3.1.*

*Proof.* Let $M = \sqrt{p} \cdot \|A^T A\|_2 + \|A^T b\|_2$. For all $x \in S$, it holds that $\|A^T(Ax - b)\|_2 \leq \|A^T A\|_2 \cdot \|x\|_2 + \|A^T b\|_2 \leq M$. First, let $x$ denote the optimal solution of $P(\lambda)$. If $x \notin \operatorname{int}(S)$,

then there exists $i \in [p]$ such that $x_i = 0$, or $\mathbf{1}_p^T x = 1$. If $\mathbf{1}_p^T x = 1$, without loss of generality we assume $x_1 > 0$. Let $x' = x - \delta \cdot e_1$ for some $\delta > 0$. Since $F_\lambda(\cdot)$ are convex, we have

$$\begin{aligned} F_\lambda(x) - F_\lambda(x') \geq (x - x')^T \nabla F_\lambda(x') &= \delta((a_1^T(Ax' - b)) + \lambda(\ln(x_1 - \delta) - \ln(1 - \delta))) \\ &\geq \delta(-M + \lambda(\ln(x_1 - \delta) - \ln(1 - \delta))). \end{aligned}$$

When $\delta \to 0^+$, we have $-\ln(1-\delta) \to +\infty$ and therefore $F_\lambda(x) - F_\lambda(x') > 0$, which contradicts with $x$ is the optimal solution of $P(\lambda)$. Otherwise, if $\mathbf{1}_p^T x < 1$, without loss of generality we assume $x_1 = 0$. Let $x' = x + \delta \cdot e_1$ for some $\delta > 0$. Since $F_\lambda(\cdot)$ are convex, we have

$$\begin{aligned} F_\lambda(x) - F_\lambda(x') \geq (x - x')^T \nabla F_\lambda(x') &= \delta((a_1^T(Ax' - b)) + \lambda(\ln(\delta) - \ln(1 - \mathbf{1}_p^T x - \delta))) \\ &\geq \delta(-M + \lambda(\ln(\delta) - \ln(1 - \mathbf{1}_p^T x - \delta))). \end{aligned}$$

When $\delta \to 0^+$, we have $-\ln(\delta) \to +\infty$ and therefore $F_\lambda(x) - F_\lambda(x') > 0$, which contradicts with $x$ is the optimal solution of $P(\lambda)$.

Then, let pair $(y, \lambda)$ satisfies that $y \in \mathrm{int}(S)$ and $\lambda \geq \lambda_{\min}$. We want to show that $y - h \cdot v(y, \lambda) \in \mathrm{int}(S)$ when $h \leq \frac{\lambda_{\min}(\mu + \lambda_{\min})}{4LG}$. Let $y_{(i)}$ be the $i$-th component of $y$, we have

$$\nabla^2 \Omega(y) = \mathrm{diag}(y_{(i)}^{-1}) + (1 - \mathbf{1}_p^T y)^{-1} \mathbf{1}_{p \times p}, \quad \left(\nabla^2 \Omega(y)\right)^{-1} = \mathrm{diag}(y_{(i)}) - yy^T.$$

Let $z = (A^T A + \lambda \nabla^2 \Omega(y))^{-1}(A^T(Ay - b))$ and $y' = y - h \cdot z$. Since $y \in S$, we have $\|A^T(Ay - b)\| \leq M_1$, and therefore,

$$\|z\| \leq \left\|(A^T A + \lambda \nabla^2 \Omega(y))^{-1}\right\| \cdot \|A^T(Ay - b)\| \leq \frac{G}{\mu + \lambda_{\min}}.$$

Let $w = \nabla^2 \Omega(y) \cdot z$, and it holds that

$$\begin{aligned} \|w\| = \|\nabla^2 \Omega(y) \cdot z\| &= \|A^T(Ay - b) - A^T A z\|_\infty \leq \|A^T(Ay - b)\| + \|A^T A\| \cdot \|z\| \\ &\leq G + \frac{LG}{\mu + \lambda_{\min}} < \frac{2LG}{\mu + \lambda_{\min}}. \end{aligned}$$

Therefore, for all $i = 1, \ldots, p$, it holds that

$$\begin{aligned} |z_{(i)}| = \frac{1}{\lambda} \cdot |y_{(i)}(w_{(i)} - y^T w)| &\leq \frac{y_{(i)}}{\lambda} \cdot (|w_{(i)}| + |y^T w|) \leq \frac{y_{(i)}}{\lambda} \cdot (\|w\|_\infty + \|w\|_\infty) \\ &< y_{(i)} \cdot \frac{4LG}{\lambda_{\min}(\mu + \lambda_{\min})}. \end{aligned}$$

Hence, when $h \leq \frac{\lambda_{\min}(\mu + \lambda_{\min})}{4LG}$, we have

$$y'_{(i)} = y_{(i)} - z_{(i)} = y_{(i)} \cdot \left(1 - h \cdot \frac{4LG}{\lambda_{\min}(\mu + \lambda_{\min})}\right) > 0.$$

Also, it holds that

$$|\mathbf{1}_p^T z| = \left| \sum_{i=1}^{p} (y_{(i)}(w_{(i)} - y^T w) \right| = (1 - \mathbf{1}_p^T y)|y^T w| \leq (1 - \mathbf{1}_p^T y) \cdot \|w\|_\infty.$$

and therefore, when $h \leq \frac{\lambda_{\min}(\mu+\lambda_{\min})}{4LG}$, we have

$$1 - \mathbf{1}_p^T y' = 1 - \mathbf{1}_p^T y + h \cdot \mathbf{1}_p^T z > (1 - \mathbf{1}_p^T y) \cdot \left(1 - h \cdot \frac{2LG}{\mu + \lambda_{\min}}\right) > 0.$$

Then we conclude that $y' \in \text{int}(S')$. Moreover, since linear interpolation is a convex combination of grid points, the approximate path $\hat{y}(\lambda)$ for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ is a subset of $\text{int}(S')$. □