Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Kernel mirror prox and RKHS gradient flow for mixed functional Nash equilibrium

Pavel Dvurechensky, Jia-Jie Zhu

submitted: July 12, 2023

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: pavel.dvurechensky@wias-berlin.de
jia-jie.zhu@wias-berlin.de

No. 3032
Berlin 2023

2010 *Mathematics Subject Classification.* 90C90, 68Q25, 65K15, 90C33.

Key words and phrases. Gradient flow, distributionally robust optimization, Nash equilibrium, optimization, kernel methods, generative models.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Kernel mirror prox and RKHS gradient flow for mixed functional Nash equilibrium

Pavel Dvurechensky, Jia-Jie Zhu

Abstract

The theoretical analysis of machine learning algorithms, such as deep generative modeling, motivates multiple recent works on the Mixed Nash Equilibrium (MNE) problem. Different from MNE, this paper formulates the Mixed Functional Nash Equilibrium (MFNE), which replaces one of the measure optimization problems with optimization over a class of dual functions, e.g., the reproducing kernel Hilbert space (RKHS) in the case of Mixed Kernel Nash Equilibrium (MKNE). We show that our MFNE and MKNE framework form the backbones that govern several existing machine learning algorithms, such as implicit generative models, distributionally robust optimization (DRO), and Wasserstein barycenters. To model the infinite-dimensional continuous-limit optimization dynamics, we propose the Interacting Wasserstein-Kernel Gradient Flow, which includes the RKHS flow that is much less common than the Wasserstein gradient flow but enjoys a much simpler convexity structure. Time-discretizing this gradient flow, we propose a primal-dual kernel mirror prox algorithm, which alternates between a dual step in the RKHS, and a primal step in the space of probability measures. We then provide the first unified convergence analysis of our algorithm for this class of MKNE problems, which establishes a convergence rate of $O(1/N)$ in the deterministic case and $O(1/\sqrt{N})$ in the stochastic case. As a case study, we apply our analysis to DRO, providing the first primal-dual convergence analysis for DRO with probability-metric constraints.

1 Introduction

Training state-of-the-art large-scale machine learning models typically requires stochastic optimization with non-convex objective functions, which has achieved great empirical success. However, its reliability and computational complexity must also be theoretically analyzed and understood. For example, the potential of deep generative models is remarkable in numerous machine learning applications such as image generation [21], reinforcement learning [24], and molecular dynamics [39]. However, training generative models effectively has been a challenging topic for machine learning research that attracts both empirical and theoretical research interests.

State-of-the-art theoretical analysis of generative models formulates the training process as finding a mixed Nash Equilibrium (MNE) in a two-player zero-sum game [25, 15, 52, 51]. The MNE problem seeks the solution, i.e., MNE of the saddle-point optimization problem

$$\inf_{\mu \in \mathcal{M}} \sup_{\nu \in \mathcal{M}} F(\mu, \nu), \tag{1}$$

where $F(\mu, \nu)$ is a bi-variate objective function(al) of the probability measures μ, ν.

In machine learning, there exists a commonly-used alternative paradigm to directly optimizing the measures, such as in the MNE (1). This paradigm exploits the duality between probability measure

DOI 10.20347/WIAS.PREPRINT.3032

space $\mathcal{M}$ and a dual functional space $\mathcal{F}$, i.e., instead of optimizing w.r.t. a measure, we solve the *functional optimization problem*

$$\inf_{f\in\mathcal{F}} \mathcal{E}(f). \tag{2}$$

Such methods often leverage scalable learning models such as reproducing kernel Hilbert spaces and deep neural networks for parameterizing and manipulating the dual functions, instead of directly searching in the space of probability distributions as in (1). Such dual functional approaches have witnessed success in many domains of machine learning research, e.g., generative modeling [40, 5, 23, 28] computing optimal transport [20], distributionally robust optimization [54], Wasserstein barycenters [31, 50, 29]. However, optimization over those functional spaces, e.g., RKHSs or DNNs, is inherently an infinite dimensional problem, whose convergence analysis does not directly follow from the finite-dimensional setting.

As an concrete application of our functional optimization framework, we formulate the *Mixed Functional Nash Equilibrium* (MFNE), whose inner optimization is an optimization problem over a set $\mathcal{F}$ of functions

$$\inf_{f\in\mathcal{F}} \sup_{\mu\in\mathcal{M}} F(\mu, f), \tag{3}$$

which, like the MNE, is a special case of the infinite-dimensional pure NE problem. MFNE (3) has appeared in several cutting-edge algorithms in ML research, which we discuss in detail in Section 2.1. In such cases of saddle-point optimization problems, a unified convergence analysis for those applications to the primal-dual setting is still missing. To fill this gap, this paper provides the convergence analysis for optimization problems that move in a functional space.

Centered around the MFNE (3), this paper makes the following technical contributions:

1 We model the infinite-dimensional continuous-time optimization dynamics of the general functional optimization problem (2) as RKHS gradient flows. Despite its simple structure, the RKHS gradient flow is less explored in machine learning but enjoys simpler structures and straightforward convexity instead of (generalized) geodesic convexity needed for optimization in the Wasserstein space.

2 As a non-trivial application, we model the MFNE (3) as an Interacting Wasserstein-Kernel Gradient Flow (19)-(14), , which couples the now-well-known Wasserstein gradient flow and an RKHS gradient flow. We show that the time-discretization of this interacting gradient flow results in a discrete-time primal-dual kernel mirror prox algorithm for solving MFNE (3).

3 The techniques of optimizing dual (to probability measures) functions using modern learning models, such as RKHSs and DNNs, have been applied in many recent works. However, a unified convergence analysis is still missing. This paper provides a unified convergence analysis for functional optimization problems in the context of MFNE (3), with a convergence rate $\mathcal{O}(\frac{1}{N})$ via a primal-dual kernel mirror prox algorithm. To the best of our knowledge, it is the first analysis with the kernel mirror prox steps in the dual functional space, which differs from the typical mean-field analysis of measure optimization.

4 As a case study, we apply our analysis to distributionally robust optimization (DRO) to establish a convergence rate via primal-dual stochastic kernel mirror prox in the case of maximum mean discrepancy ambiguity sets. To the best of our knowledge, this is the first primal-dual continuous optimization guarantee for DRO.

5 Last but not least, the unification perspective of a few learning tasks, provided by MFNE and MKNE, highlights our Kernel Mirror Prox as a general-purpose functional optimization algorithm for optimizing over probability measures in the dual space, with theoretical convergence guarantee. This is similar to general-purpose algorithms such as Langevin Monte-Carlo, Stein variational gradient descent.

2 Preliminaries

Notation We use $\mathcal{M} = \mathrm{Prob}(\bar{\Omega})$ to denote the space of probability measures defined on the closure of a closed bounded convex domain $\Omega \subset \mathbb{R}^d$. We say that (3) is convex-concave if the inner problem is concave maximization and the outer problem is convex minimization. The convexity notion, if not otherwise specified, refers to the regular convexity notion (defined in (13)). For PDE gradient flows, the states are functions of both time and space, for example, $u(t, x)$. When there is no ambiguity, we write $u(t) := u(t, \cdot)$ to denote the function at evolutionary time t. If not otherwise specified, proofs are deferred to the appendix.

2.1 Duality of metrics on probability measures

One reason behind the ubiquity of the functional optimization problem (2) and MFNE (3) in machine learning problems is that common probability metrics admit a dual characterization. For example, the *optimal transport* distance [45, 2], e.g., p-Wasserstein metric W_p, can be characterized in the dual space via the dual Kantorovich problem

$$\mathrm{OT}(\mu, \nu) = \sup_{\psi_1, \psi_2} \int \psi_1 \,\mathrm{d}\mu + \int \psi_2 \,\mathrm{d}\nu \tag{4}$$

$$\text{s.t. } \psi_1(x) + \psi_2(y) \leq c(x, y),\ \forall x, y,\ \text{a.e.} \tag{5}$$

which is an infinite-dimensional optimization problem with an infinite constraint. Bounded continuous functions ψ_i are referred to as the Kantorovich potential functions. $c(x, y)$ is the transport cost function associated with the transport. In the machine learning literature, researchers have explored this dual formulation by directly parameterizing the Kantorovich potential, e.g., using RKHS functions [20] or ICNNs [31, 28].

Another commonly used metric is the integral probability metric (IPM), which is defined via the weak norm formulation

$$\mathrm{IPM}(\mu, \nu) = \sup_{f \in \mathcal{F}} \int f \mathrm{d}(\mu - \nu).$$

One particular choice of the test function family is the RKHS-norm-ball $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$, which yields the kernel maximum mean discrepancy (MMD) [22]. $\mathcal{F} = \{f : \mathrm{Lip}\,(f) \leq 1\}$ recovers the type-1 (Kantorovich) Wasserstein metric.

Those basic duality results characterize the relationship of the underlying geometry of the learning problem and the variational problem that optimizes w.r.t. functions, such as the IPM test functions and the Kantorovich potentials in OT. This has been exploited in several fields in machine learning, which we detail below.

Implicit generative models (IGM) We consider the following IGM formulation, such as generative moment matching networks [32, 17] and generative adversarial networks (GAN) [21],

$$\inf_{G_\theta} \mathbb{E}_Z \mathcal{D}(P, G_\theta(Z)), \tag{6}$$

where $\mathcal{D}$ can be chosen as a discrepancy measure such as the f-divergence family, optimal transport distance, or kernel maximum mean discrepancy (MMD). In IGM, P is often taken to be the training data distribution. Recent theoretical analysis of the optimization for training (6) *lifts* the non-convex optimization problem to the space of probability measures, which is an instantiation of a mixed Nash equilibrium (MNE) problem (1). The problem can then be cast into this paper's general MFNE formulation (3), for example, by choosing the function $\mathcal{D}$ as the integral probability metric (IPM) family

$$\inf_{\mu\in\mathcal{M}} \sup_{f\in\mathcal{F}} \left\{ \int f(x)dP(x) - \mathbb{E}_{\theta\sim\mu} \int f(g_\theta(z))dQ(z) \right\}, \tag{7}$$

where $g_\theta(z)$ is the generator density. If the function class $\mathcal{F}$ is chosen to be the class of 1-Lipschitz functions, then the formulation is the Wasserstein GAN [5, 23]. On the other hand, if $\mathcal{F}$ is the RKHS, this is the MMD GAN [30, 8]. Note that this lifted problem is now *convex-concave* in the optimization variables.

Distributionally robust optimization (DRO) We now consider a special case: the metric-ball-constrained DRO problem [14]

$$\inf_{\theta} \sup_{\mu:\mathcal{D}(\mu,\widehat{P}_n)\leq\epsilon} \mathbb{E}_\mu[l(\theta;x)] \tag{8}$$

Below, we consider DRO with the probability metric $\mathcal{D}$ to be the p-Wasserstein DRO ($p \geq 1$) [18] as well as the Kernel(-MMD) DRO [54]. Different from common DRO reformulation using conic linear duality, we propose a primal-dual reformulation of DRO.

Lemma 2.1 (Primal-dual reformulation of Wasserstein and kernel DRO)**.** *Suppose the probability metric is chosen to the MMD, then the DRO problem* (8) *admits the reformulation*

$$\inf_{\theta\in\mathbb{R}^d, f\in\mathcal{H}} \sup_{\mu\in\mathcal{M}} \mathbb{E}_\mu(l-f) + \frac{1}{N}\sum_{i=1}^{N} f(x_i) + \epsilon\|f\|_{\mathcal{H}}. \tag{9}$$

Furthermore, it is equivalent to the smoothed optimization problem

$$\inf_{\theta\in\mathbb{R}^d, f\in\mathcal{H}} \sup_{\mu\in\mathcal{M}, h\in\mathcal{H}:\|h\|_{\mathcal{H}}\leq 1} \left(\frac{1}{N}\sum_{i=1}^{N} f(x_i) \quad + \quad \epsilon\langle h, f\rangle \quad + \quad \mathbb{E}_\mu(l(\theta;x) - f(x)) \right). \tag{10}$$

Suppose the probability metric is chosen to the optimal transport metric, e.g., p-Wasserstein distance. Then, the DRO problem (8) *admits the reformulation*

$$\inf_{\gamma>0,\theta\in\mathbb{R}^d, f\in\Psi_c} \sup_{\mu\in\mathcal{M}} \mathbb{E}_\mu(l-\gamma\cdot f) - \frac{\gamma}{N}\sum_{i=1}^{N} f^c(x_i) + \gamma\cdot\epsilon. \tag{11}$$

Ψ_c is the set of c-concave [45] functions and $f^c(y) := \inf_x c(x,y) - f(x)$ denotes the c-transform.

Lemma 2.1 shows that those primal-dual DRO formulations have the *convex-concave* structure in μ, f as in the MFKE, and are convex in the learning model θ when the loss l is. (10) is concave in the smoothing variable h. (11) is trivially convex in the dual variable γ.

Wasserstein barycenter The Wasserstein barycenter problem [1, 31, 50, 29] can be formulated as a saddle-point optimization problem.

$$\min_{\mu\in\mathcal{M}}\sum_{i=1}^{n}\alpha_i\left[\mathcal{W}(\mu,\nu_i)\right] \quad = \quad \min_{\mu\in\mathcal{M}}\sum_{i=1}^{n}\alpha_i\sup_{f_i\in\Psi_c}\left\{\int f_i^c \mathrm{d}\mu+\int f_i \mathrm{d}\nu_i\right\}, \tag{12}$$

where $\mu_i \in \mathcal{M}$ are given probability measures, $f_i \in \Psi_c$ are the Kantorovich potentials associated with the Wasserstein distance. f_i^c again denote the c-transform.

Note that in the aforementioned settings of p-Wasserstein ($p \neq 1$) metric, one may parameterize Kantorovich potential functions using tools such as the random Fourier features [44] and input convex neural networks [3], such as done in [20, 35, 28]. In our theoretical analysis, we focus on the RKHS functions (and hence the kernel MMD) setting in the rest of the paper due to the difficulty in theoretically characterizing the approximation error of deep models such as ICNNs.

2.2 Gradient Flow in the Wasserstein and Hilbert Spaces

Recent theoretical analysis of generative models via MNE, e.g., [25, 15, 52, 51], adopted the mean-field limit point of view closely related to the mathematical topic on PDE gradient flow of probability measures [2, 41, 26, 42]. Notably, works such as [25] modeled training dynamics of GAN as sampling using Langevin SDE, which is also equivalent to solving the Fokker-Planck PDE.

Intuitively, a gradient flow describes a dynamical system that is driven towards the dissipation of certain energies. This system is called a *gradient system*. For example, the dynamical system described by an ordinary differential equation in the Euclidean space that follows the negative gradient direction, $\dot{x}(t) = -\nabla f(x(t)), x(t) \in \mathbb{R}^d$, is a simple gradient system. One milestone of the gradient system research is the works of Otto and colleagues in deriving the Wasserstein gradient flow [41, 26, 42]. Rigorous characterizations of general metric gradient systems have been carried out in PDE literature, for which we refer to [2, 45] for complete treatments and [43, 36] for a first principles' introduction, whose perspective we adopt in this paper.

Recent machine learning literature has explored the *Wasserstein gradient system*, which generates the *Wasserstein gradient flow* (WGF). It describes the evolution of probability measures in the Wasserstein metric space $(\mathcal{M}, W_p)$, driven by some energy functionals. In the case of 2-Wasserstein metric, the metric space $(\mathcal{M}, W_2)$ has particularly nice properties, namely, it is a so-called geodesic metric space [2]. The gradient flow equation of Boltzmann entropy associated with this Wasserstein gradient system is precisely the diffusion equation, which coincide with the heat equation from the PDE perspective. From the optimization perspective, those foundational works allow us to view complex dynamical systems as optimization algorithms of probability measures, which minimizes or maximizes objective function(al)s (e.g., energy) in particular geometries (e.g., Wasserstein space). This is vividly described by a quote from Felix Otto:

> The merit of the right gradient flow formulation of a dissipative evolution equation is that it separates energetics and kinetics: The energetics endow the state space with a *functional*, the kinetics endow the state space with a (Riemannian) *geometry* via the metric tensor.

While WGF has sparked significant interest within the machine learning community recently, performing optimization with WGF requires extra care since its convergence cannot be characterized using the regular notion of convexity. Recall that a functional $\mathcal{E}$ defined on a Hilbert space $\mathcal{H}$ is λ-convex if $\forall s \in [0,1], \forall u_0, u_1 \in \mathcal{H}$,

$$\mathcal{E}((1 - s)u_0 + su_1) \leq (1 - s)\mathcal{E}(u_0) + s\mathcal{E}(u_1) - \frac{\lambda}{2}s(1 - s)\|u_0 - u_1\|^2_{\mathcal{H}}. \quad (13)$$

If $\lambda > 0$, $\mathcal{E}$ is strongly convex. This notion of convexity does not make sense in the Wasserstein space and one must summon the generalized geodesic convexity [2]. In contrast, gradient flows in the Hilbert space geometry enjoy a much simpler structure and stronger characterization results. We show below that the dual functional optimization problems can be characterized by a simpler gradient flow in the reproducing kernel Hilbert space, where the regular convexity notion (13) is sufficient.

3 RKHS Gradient Flow

Our starting point is to model the infinite-dimensional continuous-time optimization dynamics of the functional optimization problem (2) using the RKHS gradient flows.

$$\partial_t f + \mathcal{E}'_f(f) = 0, \quad f(0, x) = f^0(x) \in \mathcal{H}. \quad (14)$$

In principle, other function spaces in the literature can also be considered in practice, such as the random Fourier features functions, single hidden-layer neural networks [6], and ICNNs. Although we focus the theoretical analysis in the rest of the paper on the RKHS setting.

Equation (14) is a gradient flow equation in a reproducing kernel Hilbert space, which is much less explored in the machine learning community than the WGF above. For example, in [4], the squared RKHS norm was used as the driving energy functional for the Wasserstein gradient flow, rather than the dissipation geometry for the flow as in our formulation. A few other works such as [12, 9, 16, 34, 27] studied the kernelized Wasserstein gradient flow in the context of Stein geometry. Those particular cases do not exploit the simplicity of gradient flow structure in the Hilbert space. Below, we derive standard results for RKHSGF in the context of this paper for completeness. For readers unfamiliar with PDE gradient flows in the Hilbert space, we refer to [2] for complete treatment and [46, 36] for accessible introductions.

We first establish the main results for RKHSGF, namely, existence, uniqueness, and a powerful result known as the *evolutionary variational inequalities* (EVI)$_\lambda$.

Lemma 3.1 (Characterizations of RKHS gradient flow)**.** *Suppose the energy functional $\mathcal{E}$ is proper, upper semicontinuous, λ-convex for some $\lambda \in \mathbb{R}$ (i.e., either convex or concave), and has compact sublevel sets. Then for any initial condition in the RKHS $f(0, x) \in \mathcal{H}$, there exists a unique solution at time t, $f(t) \in \mathcal{H}$.*

Furthermore, the gradient flow solution $f(t, x)$ satisfies (EVI)$_\lambda$, for $t \in [0, T]$.

$$\frac{1}{2}\|f(t) - \nu\|^2_{\mathcal{H}} \leq \frac{1}{2}\mathrm{e}^{-\lambda(t-s)}\|f(s) - \nu\|^2_{\mathcal{H}} + M_\lambda(t - s)(\mathcal{E}(\nu) - \mathcal{E}(f(t))),$$
$$M_\lambda(\tau) = \int_0^\tau e^{-\lambda(\tau - s)}\mathrm{d}s, \quad \forall \nu \in \mathrm{dom}(\mathcal{F}) \subset \mathcal{H}.$$

Using (EVI)$_\lambda$, we can effortlessly extract convergence results. Suppose a minimizer of the energy exists $f^* \in \inf_{f\in\mathcal{H}} \mathcal{E}(f)$, we set $\nu = f^*, s = 0$ in (EVI)$_\lambda$

$$\begin{aligned}\|f(t) - f^*\|^2_{\mathcal{H}} \leq \mathrm{e}^{-\lambda t}\|f(0) - f^*\|^2_{\mathcal{H}} \\ + 2M_\lambda(t-s)\left(\inf_{f\in\mathcal{H}} \mathcal{E}(f) - \mathcal{E}(f(t))\right) \\ \leq \mathrm{e}^{-\lambda t}\|f(0) - f^*\|^2_{\mathcal{H}}, \quad (15)\end{aligned}$$

yielding an *exponential convergence in time* if the energy is convex (in the usual sense) w.r.t. the f variable, i.e., $\lambda > 0$. Note that the convexity condition can be further weakened using functional inequalities such as the logarithmic Sobolev inequality.

One important distinction between the functional optimization dynamics in RKHSGF and the measure optimization in MNE optimization dynamics is that we do not need advanced structures such as *generalized geodesic convexity* from the WGF setting — linearity or regular convexity is sufficient for RKHSGF. This can be seen in the following example.

Example 3.1. [Difference in convexity] In a 2-Wasserstein space $(\mathcal{M}, W_2)$, linear energy functional $\mathcal{E}_0(\mu) = \int V_0 \mathrm{d}\mu$ for $\mu \in (\mathcal{M}, W_2)$, is non-convex geodesically if and only if the function $V_0(x)$ is non-convex [2].

On the other hand, in an RKHS $\mathcal{H}$, linear energy functional $\mathcal{E}_1(f) = \langle V_1, f\rangle_{\mathcal{H}}$ for $f \in \mathcal{H}$ is always convex, regardless of the non-convexity of V_1.

Hence, while the MNE problem "lifts"the non-convex optimization problems to the measure spaces, the resulting objective functions as in WGF are not *geodesically convex*. Hence, the resulting WGF does not necessarily converge globally even in the linear case above. In other words, there is still no free lunch despite the lifting to the linear structure. In contrast, the RKHSGF in our formulation converges under the usual convexity in the Hilbert space. This distinction of convexity has also been exploited in the context of distributionally robust optimization with nonlinear or nonconvex (in the uncertain variable) DRO objective functions using kernel methods [54]. The same does not hold true for the Wasserstein DRO, which requires Lagrangian relaxation under nonconvex objectives [48].

The standard proof of Lemma 3.1 is via time-discretization using the following *minimizing movement scheme* (MMS), also known as the time-incremental minimization scheme

$$\hat{f}^{k+1} \in \arg\inf_{f\in\mathcal{H}} \left\{\mathcal{E}(f) + \frac{1}{2\tau}\|f - \hat{f}^k\|^2_{\mathcal{H}}\right\}, \tag{16}$$

$$\hat{f}^0(x) = f(0,x) \in \mathcal{H}, \tag{17}$$

where $\tau > 0$ is the step size for time-discretization. Defining the piecewise constant function $\bar{f}_\tau(t,x) := \hat{f}^k(x)$ for $t \in [k\tau, k\tau + \tau]$, standard PDE proofs (see, e.g., [2]) guarantee that $\bar{f}_\tau$ converges to the continuous-time RKHSGF solution, i.e., $\bar{f}_\tau \to f(t) \text{ as } \tau \to 0$.

Therefore, a natural bi-product of the existence results for RKHSGF is the MMS step (16). This is a fully-implicit Euler discretization of the RKHSGF, which is difficult to implement in practice. In the next section, we will use the explicit step, coupled with measure-update step, to derive the *primal-dual kernel mirror prox* algorithm, which is the discrete-time counterpart to the IWKGF (19)-(14).

Example: Interacting Wasserstein-Kernel Gradient Flow Previously, the authors of [15] proposed the Interacting Wasserstein Gradient Flow as the infinite-dimensional continuous-time optimization dynamics for solving the MNE problem (1). We now propose a new coupled gradient system which alternatives between a WGF and an *RKHS gradient flow* (RKHSGF). We term the mean-field dynamics *Interacting Wasserstein-Kernel Gradient Flow* (IWKGF). Our perspective is to choose the (dual) functional space in MFNE (3) as a reproducing kernel Hilbert space (RKHS) [53, 49, 7, 47]. That is to say, let $\mathcal{F} = \mathcal{H}$ in MFNE (3), we solve a specific variation of MFNE, which is the *Mixed Kernel Nash Equilibrium* (MKNE)

$$\inf_{f\in\mathcal{H}} \sup_{\mu\in\mathcal{M}} F(\mu, f). \tag{18}$$

The gradient flow equations govern the IWKGF system are

$$\partial_t \mu - \nabla \cdot (\mu \cdot F'_\mu(\mu, f)) = 0, \quad \mu(0, x) = \mu^0(x) \in \mathcal{M}, \tag{19}$$

$$\partial_t f + F'_f(\mu, f) = 0, \quad f(0, x) = f^0(x) \in \mathcal{H}. \tag{20}$$

where $\mathcal{H}$ is an RKHS. The derivatives F'_μ, F'_f are taken in the sense of the Fréchet differential. We will calculate the concrete forms in the next sections.

When viewed standalone, The properties of the RKHS gradient flow (20) has already been discussed above. The Fokker-Planck equation (19) can be viewed either as a Wasserstein gradient flow of the energy functional $\mathcal{E}_0(\mu) := F(\mu, f)$ from the PDE perspective, or as a Langevin SDE, and is already well understood in the machine learning community. We refer to [46, 2] for its properties and recent works such as [38, 11, 10] for recent machine learning applications.

4 A Primal-Dual Kernel Mirror Prox Algorithm

To construct our mirror prox algorithm, we consider generic variable $x \in \mathbb{R}^p$ with domain $\mathcal{X}$ and make some mild regularity assumptions. Namely, we restrict $\mathcal{M}$ to the set of probability measures on $\mathcal{X}$ that admit densities w.r.t. the Lebesgue measure and have a density that is continuous and positive almost everywhere on $\mathcal{X}$. We also assume that there is a Hilbert space $\mathcal{H}$ and a convex and closed set $H \in \mathcal{H}$. For the sake of generality, we consider a slight variation of MKNE (18) as the following general infinite-dimensional saddle-point problem on the spaces of measures and functions

$$\inf_{f\in H\subseteq\mathcal{H}} \sup_{\mu\in\mathcal{M}} \quad F(f, \mu). \tag{21}$$

For shortness, we denote the set of all variables by $u = (f, \mu)$. We consider here the setting of two variables only for simplicity. An extension for a more general problem formulation covering the DRO problem (10) is easy to derive. Moreover, in the next section, we consider DRO problem (10) as a particular case study and provide technical details to check that the main assumptions of this section hold for that problem. Our first main assumption in this section is as follows.

Assumption 4.1. The functional $F(f, \mu)$ is convex in f for fixed μ and concave in μ for fixed f.

4.1 Preliminaries

To construct the mirror prox algorithm for problem (21) we need, first, to introduce proximal setup, which consists of norms, their dual, and Bregman divergences on each space of the variables.

For the space of the variable f, we use the self-dual norm of the Hilbert space $\|\cdot\|_{\mathcal{H}}$, distance generating function $d_f(f) = \frac{1}{2}\|f\|^2_{\mathcal{H}}$, which gives Bregman divergence $B_{\mathcal{H}}(f, \breve{f}) = \frac{1}{2}\|f - \breve{f}\|^2_{\mathcal{H}}$. This leads to the mirror step, which is an explicit version of the MMS step (16),

$$f_+ = \mathsf{Mirr}^{f,H}_{\eta}(f, \xi_f) = \arg\min_{\tilde{f}\in H}\{\langle \tilde{f}, \eta\xi_f\rangle + \frac{1}{2}\|\tilde{f} - f\|^2_{\mathcal{H}}\}. \tag{22}$$

For the space of the variable μ, we follow [25] and, first, introduce the Total Variation norm for the elements of $\mathcal{M}$ $\|\mu\|_{TV} = \sup_{\|\xi\|_{L^\infty}\le 1}\int \xi d\mu = \sup_{\|\xi\|_{L^\infty}\le 1}\langle \xi, \mu\rangle$, where $\|\xi\|_{L^\infty}$ is the L^∞-norm of functions. To define the mirror step, we use (negative) Shannon entropy and its Fenchel dual defined respectively as

$$\Phi(\mu) = \int d\mu \ln\frac{d\mu}{dx}, \quad \Phi^*(\xi) = \ln\int e^{\xi}dx \tag{23}$$

defined for ξ from the space $\mathcal{F}$ of all bounded integrable functions on $\mathcal{X}$. The corresponding Bregman divergence is the relative entropy given by

$$D_{\Phi}(\mu, \breve{\mu}) = \int d\mu \ln\frac{d\mu}{d\breve{\mu}}. \tag{24}$$

This leads to the mirror step [25] [Theorem 1]

$$\mu_+ = \mathsf{Mirr}^{\mu}_{\eta}(\mu, \xi_\mu) = d\Phi^*(d\Phi(\mu) - \eta\xi_\mu) \equiv d\mu_+ = \frac{e^{-\eta\xi_\mu}d\mu}{\int e^{-\eta\xi_\mu}d\mu}. \tag{25}$$

Our second main assumption is as follows.

Assumption 4.2. The functional $F(f, \mu)$ is Fréchet differentiable w.r.t. each variable and the derivatives are Lipschitz continuous in the following sense

$$\|F'_f(u) - F'_f(\tilde{u})\|_{\mathcal{H}} \le L_{ff}\|f - \tilde{f}\|_{\mathcal{H}} + L_{f\mu}\|\mu - \tilde{\mu}\|_{TV}, \tag{26}$$

$$\|F'_\mu(u) - F'_\mu(\tilde{u})\|_{L^\infty} \le L_{\mu f}\|f - \tilde{f}\|_{\mathcal{H}} + L_{\mu\mu}\|\mu - \tilde{\mu}\|_{TV}. \tag{27}$$

We also denote

$$L = \max_{\kappa_1,\kappa_2\in\{f,\mu\}}\{L_{\kappa_1\kappa_2}\}. \tag{28}$$

4.2 Kernel Mirror Prox Algorithm and Its Analysis

The updates of the ideal general infinite-dimensional mirror prox algorithm for problem (21) are given in Algorithm 1. For the analysis of the mirror prox algorithm we need the following auxuliary results. The first one is used for the mirror steps applied to the variable f.

Lemma 4.3. *Let $\mathcal{H}$ be (possibly finite-dimensoinal) Hilbert space and let $H \subset \mathcal{H}$ be convex and closed. Let $\tilde{h} \in H$ and $\xi, \tilde{\xi} \in \mathcal{H}^* = \mathcal{H}$, and*

$$h = \arg\min_{\hat{h}\in H}\left\{\langle \hat{h}, \eta\xi\rangle + \frac{1}{2}\|\tilde{h} - \hat{h}\|^2_{\mathcal{H}}\right\} = \mathit{Mirr}^{h,H}_{\eta}(\tilde{h}, \xi), \tag{29}$$

$$\tilde{h}_+ = \arg\min_{\hat{h}\in H}\left\{\langle \hat{h}, \eta\tilde{\xi}\rangle + \frac{1}{2}\|\tilde{h} - \hat{h}\|^2_{\mathcal{H}}\right\} = \mathit{Mirr}^{h,H}_{\eta}(\tilde{h}, \tilde{\xi}). \tag{30}$$

Then, for any $\hat{h} \in H$

$$\langle h - \hat{h}, \eta\tilde{\xi}\rangle \le \frac{1}{2}\|\hat{h} - \tilde{h}\|^2_{\mathcal{H}} - \frac{1}{2}\|\hat{h} - \tilde{h}_+\|^2_{\mathcal{H}} + \frac{\eta^2}{2}\|\tilde{\xi} - \xi\|^2_{\mathcal{H}} - \frac{1}{2}\|h - \tilde{h}\|^2_{\mathcal{H}}.$$

Algorithm 1 Ideal General Mirror-Prox

```
Require: Initial guess $(\tilde{f}_0, \tilde{\mu}_0)$, step-sizes $\eta_f, \eta_\mu > 0$.
for $k = 0, 1, \dots, N-1$ do
    Compute
        $f_k = \mathrm{Mirr}^{f,H}_{\eta_f}(F'_f(\tilde{u}_k)),$
        $\mu_k = \mathrm{Mirr}^{\mu}_{\eta_\mu}(\mu_k, -F'_\mu(\tilde{u}_k)).$
    Compute
        $\tilde{f}_{k+1} = \mathrm{Mirr}^{f,H}_{\eta_f}(F'_f(u_k)),$
        $\tilde{\mu}_{k+1} = \mathrm{Mirr}^{\mu}_{\eta_\mu}(\mu_k, -F'_\mu(u_k)).$
end for
Compute $\bar{u}_N = \frac{1}{N}\sum_{k=0}^{N-1} u_k$.
```

The second result characterizes the mirror step with respect to the measure μ.

Lemma 4.4 ([25] [Lemma 5]). *Let $\tilde{\mu} \in \mathcal{M}$ and $\xi, \tilde{\xi} \in \mathcal{F}$, and*

$$\mu = \mathit{Mirr}^{\mu}_{\eta}(\tilde{\mu}, \xi), \qquad \tilde{\mu}_+ = \mathit{Mirr}^{\mu}_{\eta}(\tilde{\mu}, \tilde{\xi}). \tag{31}$$

Then, for any $\hat{\mu} \in \mathcal{M}$

$$\langle \mu - \hat{\mu}, \eta\tilde{\xi} \rangle \leq D_\Phi(\hat{\mu}, \tilde{\mu}) - D_\Phi(\hat{\mu}, \tilde{\mu}_+) + \frac{\eta^2}{8}\|\tilde{\xi} - \xi\|^2_{L^\infty} - 2\|\mu - \tilde{\mu}\|^2_{TV}. \tag{32}$$

The following result gives the convergence rate of Algorithm 1.

Theorem 4.5. *Let Assumptions 4.1, 4.2 hold. Let also the stepsizes in Algorithm 1 satisfy $\eta_f = \eta_\mu = \frac{1}{16L}$, where L is defined in* (28)*. Then, for any compact set $U = U_f \times U_\mu \subseteq H \times \mathcal{M}$, the sequence $(\bar{f}_N, \bar{\mu}_N)$ generated by Allgorithm 1 satisfies*

$$\max_{\mu \in U_\mu} F(\bar{f}_N, \mu) - \min_{f \in U_f} F(f, \bar{\mu}_N) \leq \frac{8L}{N} \max_{u \in U} \left(\|f - \tilde{f}_0\|^2_{\mathcal{H}} + 2D_\Phi(\mu, \tilde{\mu}_0) \right).$$

4.3 Analysis of Stochastic Kernel Mirror Prox

To account for potential inexactness in the first-order information, we assume that instead of exact derivatives, the algorithm uses their inexact counterparts $\tilde{F}'_f(u), \tilde{F}'_\mu(u)$ that may be random and are assumed to satisfy the following assumption.

Assumption 4.6.

$$F'_f(u) = \mathbb{E}\tilde{F}'_f(u), \quad F'_\mu(u) = \mathbb{E}\tilde{F}'_\mu(u), \tag{33}$$

$$\mathbb{E}\|F'_f(u) - \tilde{F}'_f(u)\|^2_{\mathcal{H}} \leq \sigma_f^2, \quad \mathbb{E}\|F'_\mu(u) - \tilde{F}'_\mu(u)\|^2_{L^\infty} \leq \sigma_\mu^2. \tag{34}$$

Theorem 4.7. *Let Assumptions 4.1–4.6 hold. Let also in Algorithm 1 the stochastic derivatives be used instead of the deterministic and the stepsizes satisfy $\eta_f = \eta_\mu = \frac{1}{16L}$, where L is defined in* (28)*. Then, for any compact set $U = H \times U_\mu \subseteq H \times \mathcal{M}$, the sequence $(\bar{f}_N, \bar{\mu}_N)$ generated by Algorithm 1 satisfies*

$$\mathbb{E}\left\{\max_{\mu\in U_\mu} F(\bar{f}_N,\mu) - \min_{f\in U_f} F(f,\bar{\mu}_N)\right\} \\ \leq \frac{8L}{N}\max_{u\in U}\left(\|f-\tilde{f}_0\|^2_{\mathcal{H}} + 2D_\Phi(\mu,\tilde{\mu}_0)\right) + \frac{3(\sigma_f^2+\sigma_\mu^2)}{16L}.$$

Let us denote $\sigma^2 = \sigma_f^2 + \sigma_\mu^2$. As we see, Theorem 4.7 guarantees the same convergence rate as in the exact case, but up to some vicinity which is governed by the level of noise. In most cases, the σ^2/L term can be made of the same order $1/N$ by using the mini-batching technique. Indeed, a mini-batch of size N allows us to reduce the variance from σ^2 to σ^2/N, see, e.g. [19]. Yet, we note that in this case, N iterations will require the number of samples $O(N^2)$.

An alternative would be to use the information about the diameter of the set U. Indeed, assume that

$$\max_{u\in U}\left(\|f-\tilde{f}_0\|^2_{\mathcal{H}} + 2D_\Phi(\mu,\tilde{\mu}_0)\right) \leq \Omega_U^2.$$

Fixing the number of steps N and choosing $\eta_f = \eta_\mu = \min\left\{\frac{1}{16L}, \frac{\Omega_U\sigma}{\sqrt{6N}}\right\}$, we obtain the following result

$$\mathbb{E}\left\{\max_{\mu\in U_\mu} F(\bar{f}_N,\mu) - \min_{f\in U_f} F(f,\bar{\mu}_N)\right\} \leq \max\left\{\frac{8L\Omega_U^2}{N}, \sqrt{\frac{3\sigma^2\Omega_U^2}{2N}}\right\}. \quad (35)$$

5 Case Study: Distributionally Robust Optimization

In this subsection we particularize the elements of Algorithm 1 for the specific DRO problem (8), (10). In contrast with the large number of reformulation techniques using the dual formulation, we propose the first principled *primal-dual convergence analysis* for DRO using our MFNE and MKNE framework.

Compared to problem (21) it has two additional variables $\theta \in \mathbb{R}^d$ and $h \in H \subset \mathcal{H}$. For these variables, the proximal setup is introduced in the same way as for the variable f. We choose $\mathcal{H}$ to be a reproducing kernel Hilbert space with kernel k.

Our main assumptions for problem (10) are

1 l is convex w.r.t. θ for all x.
2 $L_0 = \sup_{x,\theta}\|\nabla_\theta l(\theta;x)\|_2 < +\infty$.
3 $\nabla_\theta l(\theta;x)$ is $L(x)$-Lipschitz w.r.t. θ and $L_1 = \sup_\mu \mathbb{E}_{x\sim\mu} L(x)^2 < +\infty$.
4 $C = \sup_x k(x,x) < +\infty$.

We note that the convexity assumption w.r.t. θ is used to obtain global convergence guarantee w.r.t. θ. When it does not hold in practice, we can still execute our primal-dual kernel mirror prox for DRO. This is not possible with other existing Wasserstein or kernel DRO algorithms. Clearly, then the objective F is convex in (θ, f) for fixed (μ, h) and concave in (μ, h) for fixed (θ, f). The Frechet derivatives of F

with respect to the variables (θ, f, μ, h) are given by

$$F'_\theta = \mathbb{E}_{x\sim\mu}\nabla_\theta l(\theta; x) \tag{36}$$

$$F'_f = \int k(x, x')d\hat{\mu}(x') + \epsilon h(x) - \int k(x, x')d\mu(x') \tag{37}$$

$$= \mathbb{E}_{x\sim\hat{\mu}}k(\cdot, x) + \epsilon h(\cdot) - \mathbb{E}_{x\sim\mu}k(\cdot, x) \tag{38}$$

$$-F'_\mu = f(\cdot) - l(\theta; \cdot) \tag{39}$$

$$-F'_h = -\epsilon f(\cdot). \tag{40}$$

Since the derivative w.r.t. θ and f have the form of expectation, we can use the following stochastic counterparts. We can take a sample of X_i's from μ to construct an unbiased stochastic derivative

$$\tilde{F}'_\theta = \frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\nabla_\theta l(\theta; X_i). \tag{41}$$

Similarly, we can take a sample of X_i's from μ and $\hat{X}_i$ from $\hat{\mu}$ to construct an unbiased stochastic derivative

$$\tilde{F}'_f = \epsilon h(\cdot) + \frac{1}{N_f}\sum_{i=1}^{N_f}(k(\cdot, \hat{X}_i) + k(\cdot, X_i)). \tag{42}$$

We summarize the result of applying our theoretical analysis to the primal-dual DRO.

Corollary 5.1. *Assumptions 4.2, 4.6 hold for the smoothed DRO problem* (10)*. Consequently, the results of Theorems 4.5 and 4.7 hold for the DRO problem* (10)*.*

Therefore, we obtain $O(1/N)$ convergence rate in the deterministic case and $O(1/\sqrt{N})$ convergence rate in the stochastic case for solving DRO with kernel mirror prox.

6 Discussion

In conclusion, this paper studies the functional optimization dynamics using the continuous-time RKHS gradient flow. As a specific application, we introduce the Mixed Functional Nash Equilibrium framework that governs several learning algorithms. We model the optimization dynamics as the Interacting Wasserstein-Kernel Gradient Flow and analyze its corresponding discrete-time primal-dual kernel mirror prox algorithm. We provide the first unified convergence analysis for the MKNE problem class and the primal-dual reformulation of DRO with probability-metric constraints.

As this paper focuses on theoretical analysis, code implementation is left for future work. We also did not include the now-standard mean-field analysis using the Wasserstein gradient flow and Langevin SDE, for which we refer to recent works such as [15, 10, 38].

References

[1] M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

[3] B. Amos, L. Xu, and J. Z. Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

[4] M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum Mean Discrepancy Gradient Flow. *arXiv:1906.04370 [cs, stat]*, Dec. 2019. arXiv: 1906.04370.

[5] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[6] F. Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[7] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[8] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[9] S. Chewi, T. Le Gouic, C. Lu, T. Maunu, and P. Rigollet. Svgd as a kernelized wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems*, 33:2098–2109, 2020.

[10] L. Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1-2):487–532, 2022.

[11] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

[12] C. Chu, K. Minami, and K. Fukumizu. The equivalence between stein variational gradient descent and black-box variational inference. *arXiv preprint arXiv:2004.01822*, 2020.

[13] B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M.-F. F. Balcan, and L. Song. Scalable kernel methods via doubly stochastic gradients. *Advances in neural information processing systems*, 27, 2014.

[14] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010. Publisher: INFORMS.

[15] C. Domingo-Enrich, S. Jelassi, A. Mensch, G. Rotskoff, and J. Bruna. A mean-field analysis of two-player zero-sum games. In *Advances in Neural Information Processing Systems*, volume 33, pages 20215–20226. Curran Associates, Inc.

[16] A. Duncan, N. Nüsken, and L. Szpruch. On the geometry of stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

[17] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization.

[18] R. Gao and A. J. Kleywegt. Distributionally Robust Stochastic Optimization with Wasserstein Distance. *arXiv:1604.02199 [math]*, July 2016. arXiv: 1604.02199.

[19] A. Gasnikov, P. Dvurechensky, and Y. Nesterov. Stochastic gradient methods with inexact oracle. *Proceedings of Moscow Institute of Physics and Technology*, 8(1):41–91, 2016. In Russian, first appeared in arXiv:1411.4218.

[20] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.

[22] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. Publisher: JMLR. org.

[23] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved Training of Wasserstein GANs. *arXiv:1704.00028 [cs, stat]*, Dec. 2017. arXiv: 1704.00028.

[24] J. Ho and S. Ermon. Generative Adversarial Imitation Learning. *arXiv:1606.03476 [cs]*, June 2016. arXiv: 1606.03476.

[25] Y.-P. Hsieh, C. Liu, and V. Cevher. Finding mixed Nash equilibria of generative adversarial networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2810–2819. PMLR, 09–15 Jun 2019.

[26] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998. Publisher: SIAM.

[27] A. Korba, P.-C. Aubin-Frankowski, S. Majewski, and P. Ablin. Kernel Stein Discrepancy Descent. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5719–5730. PMLR, July 2021. ISSN: 2640-3498.

[28] A. Korotin, V. Egiazarian, A. Asadulaev, A. Safin, and E. Burnaev. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

[29] A. Korotin, L. Li, J. Solomon, and E. Burnaev. Continuous Wasserstein-2 Barycenter Estimation without Minimax Optimization. *arXiv:2102.01752 [cs, stat]*, Feb. 2021. arXiv: 2102.01752.

[30] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

[31] L. Li, A. Genevay, M. Yurochkin, and J. Solomon. Continuous Regularized Wasserstein Barycenters. *arXiv:2008.12534 [cs, stat]*, Oct. 2020. arXiv: 2008.12534.

[32] Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International conference on machine learning*, pages 1718–1727. PMLR, 2015.

[33] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

[34] Q. Liu and D. Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *arXiv:1608.04471 [cs, stat]*, Sept. 2019. arXiv: 1608.04471.

[35] A. Makkuva, A. Taghvaei, S. Oh, and J. Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

[36] A. Mielke. An introduction to the analysis of gradients systems, June 2023. arXiv:2306.05026 [math-ph].

[37] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, Jan. 2009.

[38] A. Nitanda, D. Wu, and T. Suzuki. Convex Analysis of the Mean Field Langevin Dynamics. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR, May 2022.

[39] F. Noé, S. Olsson, J. Köhler, and H. Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, Sept. 2019.

[40] S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[41] F. Otto. *Double degenerate diffusion equations as steepest descent*. Citeseer.

[42] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. Publisher: Taylor & Francis.

[43] M. A. Peletier. Variational Modelling: Energies, gradient flows, and large deviations. *arXiv:1402.1990 [math-ph]*, Feb. 2014. arXiv: 1402.1990.

[44] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[45] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015. Publisher: Springer.

[46] F. Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, Apr. 2017.

[47] B. Schölkopf, A. J. Smola, F. Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[48] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. *arXiv:1710.10571 [cs, stat]*, May 2020. arXiv: 1710.10571.

[49] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[50] D. Tiapkin, A. Gasnikov, and P. Dvurechensky. Stochastic Saddle-Point Optimization for Wasserstein Barycenters. *arXiv:2006.06763 [cs, math, stat]*, Dec. 2021. arXiv: 2006.06763.

[51] C. G. Trillos and N. G. Trillos. On adversarial robustness and the use of wasserstein ascent-descent dynamics to enforce it. *arXiv preprint arXiv:2301.03662*, 2023.

[52] G. Wang and L. Chizat. An exponentially converging particle method for the mixed nash equilibrium of continuous games. *arXiv preprint arXiv:2211.01280*, 2022.

[53] H. Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[54] J.-J. Zhu, W. Jitkrittum, M. Diehl, and B. Schölkopf. Kernel Distributionally Robust Optimization: Generalized Duality Theorem and Stochastic Approximation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 280–288. PMLR, Mar. 2021. ISSN: 2640-3498.

A Further Technical Background

A.1 Proof of Lemma 2.1

We now prove Lemma 2.1, of which a more detailed version is stated below.

Lemma A.1 (Primal-dual reformulation of Wasserstein and kernel DRO). *Suppose the probability metric is chosen to the MMD, then the DRO problem* (8) *admits the following equivalent primla-dual reformulations*

$$\inf_{\theta\in\mathbb{R}^d, f\in\mathcal{H}, \tau>0} \sup_{\mu\in\mathcal{M}} \mathbb{E}_\mu(l(\theta;x)-f(x)) + \frac{1}{N}\sum_{i=1}^N f(x_i) + \frac{\tau}{2}\|f\|_{\mathcal{H}}^2 + \frac{\epsilon^2}{2\tau}, \tag{43}$$

$$\inf_{\theta\in\mathbb{R}^d, f\in\mathcal{H}} \sup_{\mu\in\mathcal{M}} \mathbb{E}_\mu(l(\theta;x)-f(x)) + \frac{1}{N}\sum_{i=1}^N f(x_i) + \epsilon\|f\|_{\mathcal{H}}. \tag{44}$$

Furthermore, it is equivalent to the smoothed optimization problem

$$\inf_{\theta\in\mathbb{R}^d, f\in\mathcal{H}} \sup_{\mu\in\mathcal{M}, h\in\mathcal{H}:\|h\|_{\mathcal{H}}\le 1} \left\{\frac{1}{N}\sum_{i=1}^N f(x_i) + \epsilon\langle h, f\rangle + \mathbb{E}_\mu(l(\theta;x)-f(x))\right\}. \tag{45}$$

Suppose the probability metric is chosen to the optimal transport metric, e.g., p-Wasserstein distance. Then, the DRO problem (8) *admits the following equivalent reformulations*

$$\inf_{\gamma>0, \theta\in\mathbb{R}^d, f\in\Psi_{\gamma\cdot c}} \sup_{\mu\in\mathcal{M}} \mathbb{E}_\mu(l(\theta;x)-f(x)) - \frac{1}{N}\sum_{i=1}^N f^{\gamma\cdot c}(x_i) + \gamma\cdot\epsilon, \tag{46}$$

$$\inf_{\gamma>0, \theta\in\mathbb{R}^d, f\in\Psi_c} \sup_{\mu\in\mathcal{M}} \mathbb{E}_\mu(l(\theta;x)-\gamma\cdot f(x)) - \frac{1}{N}\sum_{i=1}^N \gamma\cdot f^c(x_i) + \gamma\cdot\epsilon. \tag{47}$$

Ψ_c denotes the set of c-concave [45] functions and $f^c(y) := \inf_x c(x,y) - f(x)$ denotes the c-transform.

Proof. For this proof, it suffices to consider the case where θ is fixed, since only the inner maximization is reformulated. We first prove the result for the MMD setting.

MMD setting. We consider the kernel mean embedding map as a linear constraint

$$\int \phi(x)\mathrm{d}\mu = h, \tag{48}$$

where h is a function in $\mathcal{H}$. By straightforward Lagrange duality and associating the linear constraint with the multiplier in the dual space, $f\in\mathcal{H}$,

$$\inf_{\tau>0, f\in\mathcal{H}} \sup_{\mu} \left\{\langle l,\mu\rangle - \frac{1}{2\tau}\|h-\hat{h}\|_{\mathcal{H}}^2 + \frac{\epsilon^2}{2\tau} - \langle f, h\rangle_{\mathcal{H}} + \int\langle f,\phi(x)\rangle_{\mathcal{H}}\mathrm{d}\mu\right\} \tag{49}$$

where $\hat{h}$ is the kernel mean embedding of the empirical measure $\hat{P}_N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$. Carrying out the quadratic optimization problem in closed form w.r.t. h and rearranging the terms, we obtain the result in (43). An optimal choice of the dual variable τ yields the equivalent reformulation (44). Smoothing via the definition of the dual norm in the Hilbert spaces, we obtain (45).

Wasserstein setting. The reformulation is a direct consequence of combining the dual Kantorovich representation of OT (4) and Lagrange duality. □

A.2 Lemma (3.1)

Since an RKHS is a Hilbert space, Lemma (3.1) is simply the (EVI)$_\lambda$ in a Hilbert space, whose proof is standard [2, 45, 36].

A.3 Technial details on geodesic convexity in the Wasserstein space

Recent machine learning literature has explored *Wasserstein gradient flow* (WGF). The 2-Wasserstein space $(\mathcal{M}, W_2)$ is geodesically convex, as defined below.

In a metric space $(\mathcal{M}, \mathcal{D})$ a curve $\gamma : [0,1] \to \mathcal{M}$ is a (constant speed) geodesic if

$$\forall r, s \in [0,1] : \quad \mathcal{D}(\gamma(r), \gamma(s)) = |s-r| \mathcal{D}(\gamma(0), \gamma(1)).$$

We refer to that as the geodesic γ connects the points $\gamma(0)$ and $\gamma(1)$ and write Geod $(\gamma(0), \gamma(1))$ for the set of all such geodesics.

Definition A.2 (Geodesic metric spaces)**.** The metric space $(\mathcal{M}, \mathcal{D})$ is a geodesic space, if for all $u_0, u_1 \in \mathcal{M}$ there exists a geodesic connecting u_0 and u_1.

Definition A.3 (Geodesic convexity)**.** A functional $\mathcal{F} : \mathcal{M} \to \mathbb{R}_\infty$ is geodesically λ-convex if $\forall u_0, u_1 \in M \exists \gamma \in \operatorname{Geod}(u_0, u_1)$,

$$\mathcal{E}(\gamma(s)) \leq (1-s)\mathcal{E}(\gamma(0)) + s\mathcal{E}(\gamma(1)) - \frac{\lambda}{2} s(1-s)\mathcal{D}(\gamma(0), \gamma(1))^2, \ \forall s \in [0,1].$$

As we have seen in the main text, e.g., discussions around Example 3.1, this complication of convexity structure makes optimization in the Wasserstein more difficult than general Hilbert spaces, which motivates our approach to work in the RKHS.

A.4 Practical Consideration about Infinite-dimensional Mirror Steps

Algorithm 1 requires two infinite-dimensional mirror steps in the functional variables μ, f. The f-update, optimization w.r.t. RKHS functions, is standard in the machine learning literature, e.g., [13, 20, 50, 54]. We now discuss the μ-update. In addition to existing works using Bregman-mirror steps, such as [25], which employ Langevin Monte-Carlo, we show a fully variational approach using Wasserstein gradient flows. While the PDE and SDE can describe the same drift-diffusion process, the resulting optimization algorithms are different. Notably, it is possible to perform deterministic optimization steps, purely relying on the Wasserstein geomtry and bypassing the discretization of Langevin SDE. The goal of the mirror step (25) is to solve the following optimization problem

$$\mu^{k+1} \in \arg \inf_{\mu \in \mathcal{M}} \int F'_\mu(\mu^k, f^k) \mathrm{d}\mu + \frac{1}{2\tau} \mathcal{D}(\mu, \mu^k). \tag{50}$$

The optimization objective is the energy of the Fokker-Planck equation, of which we can take the time-discretization in the Wasserstein space. This results in the standard *JKO steps* for Wasserstein gradient flow [41, 26, 42]. Let $\xi^0 = \mu^k$, for the sub-step count $l = 0, ..., T$, step size $s > 0$, we solve the variational problem

$$\xi^{l+1} \in \arg \inf_{\xi \in \mathcal{M}} \int F'_\mu(\mu^k, f^k) \mathrm{d}\xi + \frac{1}{2\tau}\mathcal{D}(\xi, \mu^k) + \frac{1}{2s} W_2^2(\xi, \xi^l). \tag{51}$$

Each step can be realized, e.g., via Stein variational gradient descent [33].

B Theoretical Analysis of Primal-Dual Kernel Mirror Descent in Section 4

For the sake of generality, in particular, to cover the DRO problem (10), we consider a more general saddle-point problem. Theorems 4.5 and 4.7 are obtained as corollaries of the results obtained in this section. To make this section self-contained and for the reader's convenience we repeat some definitions and results given in the main text.

We consider generic variable $x \in \mathbb{R}^p$ with domain $\mathcal{X}$. We denote by $\mathcal{M}$ the set of all probability measures on $\mathcal{X}$ that admit densities w.r.t. the Lebesgue measure and the density is continuous and positive almost everywhere on $\mathcal{X}$. We also assume that there are two Hilbert spaces $\mathcal{H}_f, \mathcal{H}_h$ and convex set $\Theta \subseteq \mathbb{R}^d$ and convex compact $H \in \mathcal{H}_h$. We consider the following general infinite-dimensional saddle-point problem

$$\inf_{\theta \in \Theta \subseteq \mathbb{R}^d, f(x) \in \mathcal{H}_f} \sup_{\mu \in \mathcal{M}, h(x) \in H \subseteq \mathcal{H}_h} F(\theta, f, \mu, h). \tag{52}$$

For shortness, we denote the set of all variables by $u = (\theta, f, \mu, h)$.

Our first main assumption is as follows.

Assumption B.1. The functional $F(\theta, f, \mu, h)$ is convex in (θ, f) for fixed (μ, h) and concave in (μ, h) for fixed (θ, f).

B.1 Preliminaries

To construct the mirror prox algorithm for problem (52) we need to first introduce proximal setup, which consists of norms, their dual, and Bregman divergences on each space of the variables.

For the space of the variable θ, we introduce the standard proximal setup with the self-dual Euclidean norm $\|\cdot\|_2$, distance-generating function $d_\theta(\theta) = \frac{1}{2}\|\theta\|_2^2$, which gives Bregman divergence $B_\theta(\theta, \breve{\theta}) = \frac{1}{2}\|\theta - \breve{\theta}\|_2^2$. This leads to the mirror step defined as

$$\theta_+ = \mathsf{Mirr}_\eta^{\theta,\Theta}(\theta, \xi_\theta) = \arg\min_{\tilde{\theta} \in \Theta} \{\langle \tilde{\theta}, \eta \xi_\theta \rangle + \frac{1}{2}\|\tilde{\theta} - \theta\|_2^2\}. \tag{53}$$

We note that our choice of the Euclidean proximal setup is made for simplicity and that other standard proximal setups are possible [37].

For the space of the variable f, we use the self-dual norm of the Hilbert space $\|\cdot\|_{\mathcal{H}_f}$, distance generating function $d_f(f) = \frac{1}{2}\|f\|_{\mathcal{H}_f}^2$, which gives Bregman divergence $B_{\mathcal{H}_f}(f, \breve{f}) = \frac{1}{2}\|f - \breve{f}\|_{\mathcal{H}_f}^2$. This leads to the mirror step

$$f_+ = \mathsf{Mirr}_\eta^{f,\mathcal{H}_f}(f, \xi_f) = \arg\min_{\tilde{f}} \{\langle \tilde{f}, \eta \xi_f \rangle + \frac{1}{2}\|\tilde{f} - f\|_{\mathcal{H}_f}^2\} = f - \eta \xi_f. \tag{54}$$

For the space of the variable μ, we follow [25] and, first, introduce the Total Variation norm for the elements of $\mathcal{M}$

$$\|\mu\|_{TV} = \sup_{\|\xi\|_{L^\infty}\le 1} \int \xi d\mu = \sup_{\|\xi\|_{L^\infty}\le 1} \langle \xi, \mu \rangle,$$

where $\|\xi\|_{L^\infty}$ is the L^∞-norm of functions. To define the mirror step, we use (negative) Shannon entropy

$$\Phi(\mu) = \int d\mu \ln \frac{d\mu}{dx} \tag{55}$$

and its Fenchel dual

$$\Phi^*(\xi) = \ln \int e^{\xi} dx \tag{56}$$

defined for ξ from the space $\mathcal{F}$ of all bounded integrable functions on $\mathcal{X}$. The corresponding Bregman divergence is the relative entropy given by

$$D_\Phi(\mu, \breve{\mu}) = \int d\mu \ln \frac{d\mu}{d\breve{\mu}}. \tag{57}$$

This leads to the mirror step [25] [Theorem 1]

$$\mu_+ = \text{Mirr}_\eta^\mu(\mu, \xi_\mu) = d\Phi^*(d\Phi(\mu) - \eta\xi_\mu) \quad \equiv \quad d\mu_+ = \frac{e^{-\eta\xi_\mu} d\mu}{\int e^{-\eta\xi_\mu} d\mu}. \tag{58}$$

Finally, for the space of the variable h, we do the same as for the variable f. Namely, we use the distance generating function $d_h(h) = \frac{1}{2}\|h\|_{\mathcal{H}_h}^2$, which gives Bregman divergence $B_{\mathcal{H}_h}(h, \breve{h}) = \frac{1}{2}\|h - \breve{h}\|_{\mathcal{H}_h}^2$. This leads to the mirror step

$$h_+ = \text{Mirr}_\eta^{h,H}(h, \xi_h) = \arg\min_{\tilde{h}\in H} \{\langle \tilde{h}, \eta\xi_h \rangle + \frac{1}{2}\|\tilde{h} - h\|_{\mathcal{H}_h}^2\}. \tag{59}$$

Our second main assumption is as follows

Assumption B.2. The functional $F(\theta, f, \mu, h)$ is Fréchet differentiable w.r.t. each variable and the derivatives are Lipschitz continuous in the following sense

$$\|F'_\theta(u) - F'_\theta(\tilde{u})\|_2 \le L_{\theta\theta}\|\theta - \tilde{\theta}\|_2 + L_{\theta f}\|f - \tilde{f}\|_{\mathcal{H}_f} + L_{\theta\mu}\|\mu - \tilde{\mu}\|_{TV} + L_{\theta h}\|h - \tilde{h}\|_{\mathcal{H}_h}, \tag{60}$$

$$\|F'_f(u) - F'_f(\tilde{u})\|_{\mathcal{H}_f} \le L_{f\theta}\|\theta - \tilde{\theta}\|_2 + L_{ff}\|f - \tilde{f}\|_{\mathcal{H}_f} + L_{f\mu}\|\mu - \tilde{\mu}\|_{TV} + L_{fh}\|h - \tilde{h}\|_{\mathcal{H}_h}, \tag{61}$$

$$\|F'_\mu(u) - F'_\mu(\tilde{u})\|_{L^\infty} \le L_{\mu\theta}\|\theta - \tilde{\theta}\|_2 + L_{\mu f}\|f - \tilde{f}\|_{\mathcal{H}_f} + L_{\mu\mu}\|\mu - \tilde{\mu}\|_{TV} + L_{\mu h}\|h - \tilde{h}\|_{\mathcal{H}_h}, \tag{62}$$

$$\|F'_h(u) - F'_h(\tilde{u})\|_{\mathcal{H}_h} \le L_{h\theta}\|\theta - \tilde{\theta}\|_2 + L_{hf}\|f - \tilde{f}\|_{\mathcal{H}_f} + L_{h\mu}\|\mu - \tilde{\mu}\|_{TV} + L_{hh}\|h - \tilde{h}\|_{\mathcal{H}_h}. \tag{63}$$

We also denote

$$L = \max_{\kappa_1,\kappa_2\in\{\theta,f,\mu,h\}} \{L_{\kappa_1\kappa_2}\}. \tag{64}$$

Algorithm 2 Ideal General Mirror-Prox

```
Require: Initial guess $(\tilde{\theta}_0, \tilde{f}_0, \tilde{\mu}_0, \tilde{h}_0)$, step-sizes $\eta_\theta, \eta_f, \eta_\mu, \eta_h > 0$.
for $k = 0, 1, \dots, N-1$ do
    Compute
        $\theta_k = \mathrm{Mirr}^{\theta,\Theta}_{\eta_\theta}(\tilde{\theta}_k, F'_\theta(\tilde{u}_k)),$    $f_k = \mathrm{Mirr}^{f,\mathcal{H}_f}_{\eta_f}(F'_f(\tilde{u}_k)),$
        $\mu_k = \mathrm{Mirr}^{\mu}_{\eta_\mu}(\mu_k, -F'_\mu(\tilde{u}_k)),$    $h_k = \mathrm{Mirr}^{h,H}_{\eta_h}(h_k, -F'_h(\tilde{u}_k)).$
    Compute
        $\tilde{\theta}_{k+1} = \mathrm{Mirr}^{\theta,\Theta}_{\eta_\theta}(\tilde{\theta}_k, F'_\theta(u_k)),$    $\tilde{f}_{k+1} = \mathrm{Mirr}^{f,\mathcal{H}_f}_{\eta_f}(F'_f(u_k)),$
        $\tilde{\mu}_{k+1} = \mathrm{Mirr}^{\mu}_{\eta_\mu}(\mu_k, -F'_\mu(u_k)),$    $\tilde{h}_{k+1} = \mathrm{Mirr}^{h,H}_{\eta_h}(h_k, -F'_h(u_k)).$
end for
Compute $\bar{u}_N = \frac{1}{N}\sum_{k=0}^{N-1} u_k$.
```

B.2 Mirror Prox Algorithm and Its Analysis

The updates of the general ideal infinite-dimensional Mirror Prox algorithm for problem (52) are given in Algorithm 2.

For the analysis of the mirror prox algorithm, we need the following auxiliary results. The first one is used for the mirror steps applied to the variables θ, f, h.

Lemma B.3. *Let $\mathcal{H}$ be (possibly finite-dimensoinal) Hilbert space and let $H \subset \mathcal{H}$ be convex and closed. Let $\tilde{h} \in H$ and $\xi, \tilde{\xi} \in \mathcal{H}^* = \mathcal{H}$, and*

$$h = \arg\min_{\hat{h} \in H} \left\{ \langle \hat{h}, \eta\xi \rangle + \frac{1}{2}\|\tilde{h} - \hat{h}\|^2_{\mathcal{H}} \right\} = \mathit{Mirr}^{h,H}_{\eta}(\tilde{h}, \xi), \tag{65}$$

$$\tilde{h}_+ = \arg\min_{\hat{h} \in H} \left\{ \langle \hat{h}, \eta\tilde{\xi} \rangle + \frac{1}{2}\|\tilde{h} - \hat{h}\|^2_{\mathcal{H}} \right\} = \mathit{Mirr}^{h,H}_{\eta}(\tilde{h}, \tilde{\xi}). \tag{66}$$

Then, for any $\hat{h} \in H$

$$\langle h - \hat{h}, \eta\tilde{\xi} \rangle \le \frac{1}{2}\|\hat{h} - \tilde{h}\|^2_{\mathcal{H}} - \frac{1}{2}\|\hat{h} - \tilde{h}_+\|^2_{\mathcal{H}} + \frac{\eta^2}{2}\|\tilde{\xi} - \xi\|^2_{\mathcal{H}} - \frac{1}{2}\|h - \tilde{h}\|^2_{\mathcal{H}}. \tag{67}$$

Proof. By the optimality condition in (66), we have for all $\hat{h} \in H$

$$\langle \eta\tilde{\xi} - (\tilde{h} - \tilde{h}_+), \hat{h} - \tilde{h}_+ \rangle \ge 0. \tag{68}$$

Rearranging, we obtain

$$\langle \tilde{h}_+ - \hat{h}, \eta\tilde{\xi} \rangle \le \langle \tilde{h}_+ - \hat{h}, \tilde{h} - \tilde{h}_+ \rangle = -\frac{1}{2}\|\tilde{h}_+ - \tilde{h}\|^2_{\mathcal{H}} - \frac{1}{2}\|\hat{h} - \tilde{h}_+\|^2_{\mathcal{H}} + \frac{1}{2}\|\hat{h} - \tilde{h}\|^2_{\mathcal{H}}. \tag{69}$$

In the same way, by the optimality condition in (65), we have for all $\hat{h} \in H$, and, in particular for $\tilde{h}_+$

$$\langle \eta\xi - (\tilde{h} - h), \tilde{h}_+ - h \rangle \ge 0. \tag{70}$$

Rearranging, we obtain

$$\langle h - \tilde{h}_+, \eta\xi \rangle \leq \langle h - \tilde{h}_+, \tilde{h} - h \rangle = -\frac{1}{2}\|h - \tilde{h}\|_{\mathcal{H}}^2 - \frac{1}{2}\|\tilde{h}_+ - h\|_{\mathcal{H}}^2 + \frac{1}{2}\|\tilde{h}_+ - \tilde{h}\|_{\mathcal{H}}^2. \tag{71}$$

Combining the last inequality with (69) and using the Fenchel inequality, we obtain

$$\begin{aligned} \langle h - \hat{h}, \eta\tilde{\xi} \rangle =& \langle \tilde{h}_+ - \hat{h}, \eta\tilde{\xi} \rangle + \langle h - \tilde{h}_+, \eta\xi \rangle + \langle h - \tilde{h}_+, \eta(\tilde{\xi} - \xi) \rangle & (72) \\ & - \frac{1}{2}\|\tilde{h}_+ - \tilde{h}\|_{\mathcal{H}}^2 - \frac{1}{2}\|\hat{h} - \tilde{h}_+\|_{\mathcal{H}}^2 + \frac{1}{2}\|\hat{h} - \tilde{h}\|_{\mathcal{H}}^2 & (73) \\ & - \frac{1}{2}\|h - \tilde{h}\|_{\mathcal{H}}^2 - \frac{1}{2}\|\tilde{h}_+ - h\|_{\mathcal{H}}^2 + \frac{1}{2}\|\tilde{h}_+ - \tilde{h}\|_{\mathcal{H}}^2 & (74) \\ & + \frac{\eta^2}{2}\|\tilde{\xi} - \xi\|_{\mathcal{H}}^2 - \frac{1}{2}\|h - \tilde{h}\|_{\mathcal{H}}^2, & (75) \end{aligned}$$

which gives the result of the Lemma. □

The second result characterizes the mirror step with respect to the measure μ.

Lemma B.4 ([25] [Lemma 5]). *Let $\tilde{\mu} \in \mathcal{M}$ and $\xi, \tilde{\xi} \in \mathcal{F}$, and*

$$\mu = \mathit{Mirr}_\eta^\mu(\tilde{\mu}, \xi), \tag{76}$$

$$\tilde{\mu}_+ = \mathit{Mirr}_\eta^\mu(\tilde{\mu}, \tilde{\xi}). \tag{77}$$

Then, for any $\hat{\mu} \in \mathcal{M}$

$$\langle \mu - \hat{\mu}, \eta\tilde{\xi} \rangle \leq D_\Phi(\hat{\mu}, \tilde{\mu}) - D_\Phi(\hat{\mu}, \tilde{\mu}_+) + \frac{\eta^2}{8}\|\tilde{\xi} - \xi\|_{L^\infty}^2 - 2\|\mu - \tilde{\mu}\|_{TV}^2. \tag{78}$$

The following result gives the convergence rate of Algorithm 2.

Theorem B.5. *Let Assumptions B.1, B.2 hold. Let also the stepsizes in Algorithm 2 satisfy $\eta_\theta = \eta_f = \eta_\mu = \eta_h = \frac{1}{16L}$, where L is defined in* (64)*. Then, for any compact set $U = U_\theta \times U_f \times U_\mu \times U_h \subseteq \Theta \times \mathcal{H}_f \times \mathcal{M} \times H$, the sequence $(\bar{\theta}_N, \bar{f}_N, \bar{\mu}_N, \bar{h}_N)$ generated by Allgorithm 2 satisfies* [1]

$$\begin{aligned} &\max_{\mu \in U_\mu, h \in U_h} F(\bar{\theta}_N, \bar{f}_N, \mu, h) - \min_{\theta \in U_\theta, f \in U_f} F(\theta, f, \bar{\mu}_N, \bar{h}_N) \\ &\leq \frac{8L}{N} \max_{u \in U} \left(\|\theta - \tilde{\theta}_0\|_2^2 + \|f - \tilde{f}_0\|_{\mathcal{H}_f}^2 + 2D_\Phi(\mu, \tilde{\mu}_0) + \|h - \tilde{h}_0\|_{\mathcal{H}_h}^2 \right). \end{aligned}$$

Proof. Applying Lemma B.3 to the step in θ, we obtain for any $\theta \in \Theta$ and for $k = 0, ..., N-1$

$$\langle \theta_k - \theta, \eta_\theta F_\theta'(u_k) \rangle \leq \frac{1}{2}\|\theta - \tilde{\theta}_k\|_2^2 - \frac{1}{2}\|\theta - \tilde{\theta}_{k+1}\|_2^2 - \frac{1}{2}\|\theta_k - \tilde{\theta}_k\|_2^2 + \frac{\eta_\theta^2}{2}\|F_\theta'(u_k) - F_\theta'(\tilde{u}_k)\|_2^2. \tag{79}$$

Summing these inequalities for $k = 0, ..., N-1$, we obtain

$$\sum_{k=0}^{N-1} \langle \theta_k - \theta, F_\theta'(u_k) \rangle \leq \frac{1}{2\eta_\theta}\|\theta - \tilde{\theta}_0\|_2^2 + \frac{R_\theta}{2\eta_\theta},$$

$$R_\theta = \sum_{k=0}^{N-1} \left(-\|\theta_k - \tilde{\theta}_k\|_2^2 + \eta_\theta^2\|F_\theta'(u_k) - F_\theta'(\tilde{u}_k)\|_2^2 \right).$$

[1] There was an obvious typo in the primal-dual gap expression in the original theorem statement in the main text. We have fixed this in the following result.

In the same way, we obtain for all $f \in \mathcal{H}_f$

$$\sum_{k=0}^{N-1} \langle f_k - f, F'_f(u_k)\rangle \leq \frac{1}{2\eta_f}\|f - \tilde{f}_0\|^2_{\mathcal{H}_f} + \frac{R_f}{2\eta_f},$$
$$R_f = \sum_{k=0}^{N-1} \left(-\|f_k - \tilde{f}_k\|^2_{\mathcal{H}_f} + \eta_f^2 \|F'_f(u_k) - F'_f(\tilde{u}_k)\|^2_{\mathcal{H}_f} \right)$$

and for all $h \in H$

$$\sum_{k=0}^{N-1} \langle h_k - h, -F'_h(u_k)\rangle \leq \frac{1}{2\eta_h}\|h - \tilde{h}_0\|^2_{\mathcal{H}_h} + \frac{R_h}{2\eta_h},$$
$$R_h = \sum_{k=0}^{N-1} \left(-\|h_k - \tilde{h}_k\|^2_{\mathcal{H}_h} + \eta_h^2 \|F'_h(u_k) - F'_h(\tilde{u}_k)\|^2_{\mathcal{H}_h} \right).$$

Finally, applying Lemma B.4 to the step in μ, we obtain for any $\mu \in \mathcal{M}$

$$\sum_{k=0}^{N-1} \langle \mu_k - \mu, -F'_\mu(u_k)\rangle \leq \frac{1}{\eta_\mu} D_\Phi(\mu, \tilde{\mu}_0) + \frac{R_\mu}{2\eta_\mu},$$
$$R_\mu = \sum_{k=0}^{N-1} \left(-\|\mu_k - \tilde{\mu}_k\|^2_{TV} + \eta_\mu^2 \|F'_\mu(u_k) - F'_\mu(\tilde{u}_k)\|^2_{L^\infty} \right).$$

By convexity of F in (θ, f) and concavity of F in (μ, h), we have, for all $(\theta, f) \in \Theta \times \mathcal{H}_f$

$$\begin{aligned}
&\frac{1}{N}\sum_{k=0}^{N-1} F(u_k) - F(\theta, f, \bar{\mu}_N, \bar{h}_N) \leq \frac{1}{N}\sum_{k=0}^{N-1} (F(u_k) - F(\theta, f, \mu_k, h_k)) \\
&\leq \frac{1}{N}\sum_{k=0}^{N-1} (\langle \theta_k - \theta, F'_\theta(u_k)\rangle + \langle f_k - f, F'_f(u_k)\rangle) \\
&\leq \frac{1}{2N\eta_\theta}\|\theta - \tilde{\theta}_0\|_2^2 + \frac{R_\theta}{2N\eta_\theta} + \frac{1}{2N\eta_f}\|f - \tilde{f}_0\|^2_{\mathcal{H}_f} + \frac{R_f}{2N\eta_f}.
\end{aligned}$$

In the same way, we obtain that, for all $(\mu, h) \in \mathcal{M} \times H$

$$\begin{aligned}
&-\frac{1}{N}\sum_{k=0}^{N-1} F(u_k) + F(\bar{\theta}_N, \bar{f}_N, \mu, h) \leq \frac{1}{N}\sum_{k=0}^{N-1} (-F(u_k) + F(\theta_k, f_k, \mu, h)) \\
&\leq \frac{1}{N}\sum_{k=0}^{N-1} (\langle \mu_k - \mu, -F'_\mu(u_k)\rangle + \langle h_k - h, -F'_h(u_k)\rangle) \\
&\leq \frac{1}{N\eta_\mu} D_\Phi(\mu, \tilde{\mu}_0) + \frac{R_\mu}{2N\eta_\mu} + \frac{1}{2N\eta_h}\|h - \tilde{h}_0\|^2_{\mathcal{H}_h} + \frac{R_h}{2N\eta_h}.
\end{aligned}$$

Combining the last two bounds, we obtain that for all $\theta \in \Theta, f \in \mathcal{H}_f, \mu \in \mathcal{M}, h \in H$ it holds that

$$\begin{aligned}
&F(\bar{\theta}_N, \bar{f}_N, \mu, h) - F(\theta, f, \bar{\mu}_N, \bar{h}_N) \\
&\leq \frac{1}{2N\eta_\theta}\|\theta - \tilde{\theta}_0\|_2^2 + \frac{1}{2N\eta_f}\|f - \tilde{f}_0\|^2_{\mathcal{H}_f} + \frac{1}{N\eta_\mu} D_\Phi(\mu, \tilde{\mu}_0) + \frac{1}{2N\eta_h}\|h - \tilde{h}_0\|^2_{\mathcal{H}_h} \\
&+ \frac{1}{2N}\left(\frac{R_\theta}{\eta_\theta} + \frac{R_f}{\eta_f} + \frac{R_\mu}{\eta_\mu} + \frac{R_h}{\eta_h} \right).
\end{aligned}$$

Our next goal is to show that
$$\frac{R_\theta}{\eta_\theta}+\frac{R_f}{\eta_f}+\frac{R_\mu}{\eta_\mu}+\frac{R_h}{\eta_h}\le 0.$$
Using the Lipschitz condition in Assumption B.2, we obtain
$$\begin{aligned}R_\theta &= \sum_{k=0}^{N-1}\left(-\|\theta_k-\tilde{\theta}_k\|_2^2+\eta_\theta^2\|F'_\theta(u_k)-F'_\theta(\tilde{u}_k)\|_2^2\right)\\ &\le \sum_{k=0}^{N-1}\left(-\|\theta_k-\tilde{\theta}_k\|_2^2+4\eta_\theta^2(L_{\theta\theta}^2\|\theta_k-\tilde{\theta}_k\|_2^2+L_{\theta f}^2\|f_k-\tilde{f}_k\|_{\mathcal{H}_f}^2+L_{\theta\mu}^2\|\mu_k-\tilde{\mu}_k\|_{TV}^2+L_{\theta h}^2\|h_k-\tilde{h}_k\|_{\mathcal{H}_h}^2)\right).\end{aligned}$$
Combining this with the similar estimates for R_f, R_μ, R_h and rearranging the terms, we obtain
$$\begin{aligned}\frac{R_\theta}{\eta_\theta}+\frac{R_f}{\eta_f}+\frac{R_\mu}{\eta_\mu}+\frac{R_h}{\eta_h} \le \sum_{k=0}^{N-1}\Big(&\|\theta_k-\tilde{\theta}_k\|_2^2(-1/\eta_\theta+4(L_{\theta\theta}^2\eta_\theta+L_{f\theta}^2\eta_f+L_{\mu\theta}^2\eta_\mu+L_{h\theta}^2\eta_h))\\ &+\|f_k-\tilde{f}_k\|_{\mathcal{H}_f}^2(-1/\eta_f+4(L_{\theta f}^2\eta_\theta+L_{ff}^2\eta_f+L_{\mu f}^2\eta_\mu+L_{hf}^2\eta_h))\\ &+\|\mu_k-\tilde{\mu}_k\|_{TV}^2(-1/\eta_\mu+4(L_{\theta\mu}^2\eta_\theta+L_{f\mu}^2\eta_f+L_{\mu\mu}^2\eta_\mu+L_{h\mu}^2\eta_h))\\ &+\|h_k-\tilde{h}_k\|_{\mathcal{H}_h}^2(-1/\eta_h+4(L_{\theta h}^2\eta_\theta+L_{fh}^2\eta_f+L_{\mu h}^2\eta_\mu+L_{hh}^2\eta_h))\Big)\le 0,\end{aligned}$$

where we used that $\eta_\theta,\eta_f,\eta_\mu,\eta_h\le\frac{1}{16L}$ for L defined in (64).

Thus, we finally obtain that for any compact $U=U_\theta\times U_f\times U_\mu\times U_h\subset\Theta\times\mathcal{H}_f\times\mathcal{M}\times H$
$$\begin{aligned}&\max_{\mu\in U_\mu,h\in U_h}F(\bar{\theta}_N,\bar{f}_N,\mu,h)-\min_{\theta\in U_\theta,f\in U_f}F(\theta,f,\bar{\mu}_N,\bar{h}_N)\\ &\le\frac{8L}{N}\max_{u\in U}\left(\|\theta-\tilde{\theta}_0\|_2^2+\|f-\tilde{f}_0\|_{\mathcal{H}_f}^2+2D_\Phi(\mu,\tilde{\mu}_0)+\|h-\tilde{h}_0\|_{\mathcal{H}_h}^2\right).\end{aligned}$$

□

B.3 Analysis in the stochastic case

To account for potential inexactness in the first-order information, we assume that instead of exact derivatives, the algorithm uses their inexact counterparts $\tilde{F}'_\theta(u),\tilde{F}'_f(u),\tilde{F}'_\mu(u),\tilde{F}'_h(u)$, that may be random and are assumed to satisfy the following assumption.

Assumption B.6.
$$F'_\theta(u)=\mathbb{E}\tilde{F}'_\theta(u), \tag{80}$$
$$F'_f(u)=\mathbb{E}\tilde{F}'_f(u), \tag{81}$$
$$F'_\mu(u)=\mathbb{E}\tilde{F}'_\mu(u), \tag{82}$$
$$F'_h(u)=\mathbb{E}\tilde{F}'_h(u), \tag{83}$$

$$\mathbb{E}\|F'_\theta(u)-\tilde{F}'_\theta(u)\|_2^2\le\sigma_\theta^2, \tag{84}$$
$$\mathbb{E}\|F'_f(u)-\tilde{F}'_f(u)\|_{\mathcal{H}_f}^2\le\sigma_f^2, \tag{85}$$
$$\mathbb{E}\|F'_\mu(u)-\tilde{F}'_\mu(u)\|_{L^\infty}^2\le\sigma_\mu^2, \tag{86}$$
$$\mathbb{E}\|F'_h(u)-\tilde{F}'_h(u)\|_{\mathcal{H}_h}^2\le\sigma_h^2. \tag{87}$$

Theorem B.7. *Let Assumptions B.1–B.6 hold. Let also in Algorithm 2 the stochastic derivatives be used instead of the deterministic and the stepsizes satisfy* $\eta_\theta = \eta_f = \eta_\mu = \eta_h = \frac{1}{16L}$*, where* L *is defined in* (64)*. Then, for any compact set* $U = U_\theta \times U_f \times U_\mu \times U_h \subseteq \Theta \times \mathcal{H}_f \times \mathcal{M} \times H$*, the sequence* $(\bar\theta_N, \bar f_N, \bar\mu_N, \bar h_N)$ *generated by Allgorithm 2 satisfies*

$$\mathbb{E}\left\{\max_{\mu\in U_\mu, h\in U_h} F(\bar\theta_N, \bar f_N, \mu, h) - \min_{\theta\in U_\theta, f\in U_f} F(\theta, f, \bar\mu_N, \bar h_N)\right\}$$
$$\leq \frac{8L}{N}\max_{u\in U}\left(\|\theta-\tilde\theta_0\|_2^2 + \|f-\tilde f_0\|_{\mathcal{H}_f}^2 + 2D_\Phi(\mu,\tilde\mu_0) + \|h-\tilde h_0\|_{\mathcal{H}_h}^2\right) + \frac{3}{16L}(\sigma_\theta^2+\sigma_f^2+\sigma_\mu^2+\sigma_h^2).$$

Proof. We proceed as in the proof of Theorem B.5 changing in Algorithm 2 the exact first-order information to its inexact counterpart. In this way, we obtain the following counterpart of (79)

$$\langle\theta_k-\theta, \eta_\theta\tilde F'_\theta(u_k)\rangle \leq \frac{1}{2}\|\theta-\tilde\theta_k\|_2^2 - \frac{1}{2}\|\theta-\tilde\theta_{k+1}\|_2^2 - \frac{1}{2}\|\theta_k-\tilde\theta_k\|_2^2 + \frac{\eta_\theta^2}{2}\|\tilde F'_\theta(u_k) - \tilde F'_\theta(\tilde u_k)\|_2^2. \tag{88}$$

Using the inequality

$$\mathbb{E}\|\tilde F'_\theta(u_k)-\tilde F'_\theta(\tilde u_k)\|_2^2 \leq 3\mathbb{E}\left(\|\tilde F'_\theta(u_k)-F'_\theta(u_k)\|_2^2 + \|\tilde F'_\theta(\tilde u_k)-F'_\theta(\tilde u_k)\|_2^2 + \|F'_\theta(u_k)-F'_\theta(\tilde u_k)\|_2^2\right) \tag{89}$$

$$\overset{\text{Assumpt.}B.6}{\leq} 6\sigma_\theta^2 + 3\mathbb{E}\|F'_\theta(u_k)-F'_\theta(\tilde u_k)\|_2^2 \tag{90}$$

and taking the expectation in the previous inequality, we obtain the following counterpart of (79)

$$\langle\theta_k-\theta, \eta_\theta F'_\theta(u_k)\rangle \leq \frac{1}{2}\mathbb{E}\|\theta-\tilde\theta_k\|_2^2 - \frac{1}{2}\mathbb{E}\|\theta-\tilde\theta_{k+1}\|_2^2 - \frac{1}{2}\mathbb{E}\|\theta_k-\tilde\theta_k\|_2^2 \tag{91}$$

$$+\frac{3\eta_\theta^2}{2}\mathbb{E}\|F'_\theta(u_k)-F'_\theta(\tilde u_k)\|_2^2 + 3\eta_\theta^2\sigma_\theta^2. \tag{92}$$

Repeating the same steps as in the proof of Theorem B.5, we obtain that for any compact $U = U_\theta \times U_f \times U_\mu \times U_h \subset \Theta \times \mathcal{H}_f \times \mathcal{M} \times H$

$$\mathbb{E}\left\{\max_{\mu\in U_\mu, h\in U_h} F(\bar\theta_N, \bar f_N, \mu, h) - \min_{\theta\in U_\theta, f\in U_f} F(\theta, f, \bar\mu_N, \bar h_N)\right\}$$
$$\leq \frac{8L}{N}\max_{u\in U}\left(\|\theta-\tilde\theta_0\|_2^2 + \|f-\tilde f_0\|_{\mathcal{H}_f}^2 + 2D_\Phi(\mu,\tilde\mu_0) + \|h-\tilde h_0\|_{\mathcal{H}_h}^2\right) + \frac{3}{16L}(\sigma_\theta^2+\sigma_f^2+\sigma_\mu^2+\sigma_h^2). \tag{93}$$

□

Let us denote $\sigma^2 = \sigma_\theta^2+\sigma_f^2+\sigma_\mu^2+\sigma_h^2$. As we see, Theorem B.7 guarantees the same convergence rate as in the exact case, but up to some vicinity which is governed by the level of noise. In most cases, the σ^2/L term can be made of the same order $1/N$ by using mini-batching technique. Indeed, a mini-batch of size N allows to change the variance from σ^2 to σ^2/N. Yet, we note that in this case, N iterations will require the number of samples $O(N^2)$.

An alternative would be to use the information about the diameter of the set U. Indeed, assume that

$$\max_{u\in U}\left(\|\theta-\tilde\theta_0\|_2^2 + \|f-\tilde f_0\|_{\mathcal{H}_f}^2 + 2D_\Phi(\mu,\tilde\mu_0) + \|h-\tilde h_0\|_{\mathcal{H}_h}^2\right) \leq \Omega_U^2.$$

Then, we obtain the following counterpart of the r.h.s. of (93) substituting $\eta_\theta = \eta_f = \eta_\mu = \eta_h = \eta$

$$\frac{\Omega_U^2}{2N\eta} + 3\sigma^2\eta.$$

Fixing the number of steps N and choosing

$$\eta_\theta = \eta_f = \eta_\mu = \eta_h = \eta = \min\left\{\frac{1}{16L}, \frac{\Omega_U\sigma}{\sqrt{6N}}\right\},$$

we obtain the following result

$$\begin{aligned}&\mathbb{E}\left\{\max_{\mu\in U_\mu, h\in U_h} F(\bar\theta_N, \bar f_N, \mu, h) - \min_{\theta\in U_\theta, f\in U_f} F(\theta, f, \bar\mu_N, \bar h_N)\right\}\\ &\leq \max\left\{\frac{8L\Omega_U^2}{N}, \sqrt{\frac{3\sigma^2\Omega_U^2}{2N}}\right\}. \end{aligned} \tag{94}$$

B.4 Case Study: Distributionally Robust Optimization

In this subsection we particularize the elements of Algorithm 2 for the specific DRO problem (10). We choose $\mathcal{H}_f = \mathcal{H}_h = \mathcal{H}$ to be a reproducing kernel Hilbert space with kernel k.

Our main assumptions for this problem are

- l is convex w.r.t. θ.
- $L_0 = \sup_{x,\theta} \|\nabla_\theta l(\theta; x)\|_2 < +\infty$.
- $\nabla_\theta l(\theta; x)$ is $L(x)$-Lipschitz w.r.t. θ and $L_1 = \sup_\mu \mathbb{E}_{x\sim\mu} L(x)^2 < +\infty$.
- $C = \sup_x k(x, x) < +\infty$.

Clearly, then the objective F is convex in (θ, f) for fixed (μ, h) and concave in (μ, h) for fixed (θ, f). The Frechet derivatives of F with respect to the variables (θ, f, μ, h) are given by

$$F'_\theta = \mathbb{E}_{x\sim\mu}\nabla_\theta l(\theta; x) \tag{95}$$

$$F'_f = \int k(x, x')d\hat\mu(x') + \epsilon h(x) - \int k(x, x')d\mu(x') = \mathbb{E}_{x\sim\hat\mu} k(\cdot, x) + \epsilon h(\cdot) - \mathbb{E}_{x\sim\mu} k(\cdot, x) \tag{96}$$

$$-F'_\mu = f(\cdot) - l(\theta; \cdot) \tag{97}$$

$$-F'_h = -\epsilon f(\cdot). \tag{98}$$

Since the derivative w.r.t. θ and f have the form of expectation, we can use the following stochastic counterparts. We can take a sample of X_i's from μ to construct an unbiased stochastic derivative

$$\tilde F'_\theta = \frac{1}{N_\theta}\sum_{i=1}^{N_\theta} \nabla_\theta l(\theta; X_i). \tag{99}$$

Similarly, we can take a sample of X_i's from μ and $\hat{X}_i$ from $\hat{\mu}$ to construct an unbiased stochastic derivative

$$\tilde{F}'_f = \epsilon h(\cdot) + \frac{1}{N_f}\sum_{i=1}^{N_f}(k(\cdot,\hat{X}_i) + k(\cdot,X_i)). \tag{100}$$

The Lipschitz constants of the derivatives are estimated in the following way. The derivative F'_θ depends only on μ and θ. Thus, $L_{\theta f} = L_{\theta h} = 0$. Further, we have

$$\|\mathbb{E}_{x\sim\mu}\nabla_\theta l(\theta_1;x) - \mathbb{E}_{x\sim\mu}\nabla_\theta l(\theta_2;x)\|_2 \le \mathbb{E}_{x\sim\mu}L(x)\|\theta_1-\theta_2\|_2 \tag{101}$$

and $L_{\theta\theta} = L_1$.

$$\|\mathbb{E}_{x\sim\mu_1}\nabla_\theta l(\theta;x) - \mathbb{E}_{x\sim\mu_2}\nabla_\theta l(\theta;x)\|_2 \le L_0\|\mu_1-\mu_2\|_{TV} \tag{102}$$

and $L_{\theta\mu} = L_0$.

The derivative F'_f depends only on μ and h. Thus, $L_{ff} = L_{f\theta} = 0$. Further, we have

$$\|\mathbb{E}_{x\sim\hat{\mu}}k(\cdot,x) + \epsilon h_1(\cdot) + \mathbb{E}_{x\sim\mu_1}k(\cdot,x) - (\mathbb{E}_{x\sim\hat{\mu}}k(\cdot,x) + \epsilon h_2(\cdot) + \mathbb{E}_{x\sim\mu_2}k(\cdot,x))\|_{\mathcal{H}} \tag{103}$$

$$\le \epsilon\|h_1-h_2\|_{\mathcal{H}} + \|\mathbb{E}_{x\sim\mu_1}k(\cdot,x) - \mathbb{E}_{x\sim\mu_2}k(\cdot,x)\|_{\mathcal{H}} \le \epsilon\|h_1-h_2\|_{\mathcal{H}} + \sqrt{C}\|\mu_1-\mu_2\|_{TV}, \tag{104}$$

i.e., $L_{f\mu} = \sqrt{C}, L_{fh} = \epsilon$. Here we used that

$$\|\mathbb{E}_{x\sim\mu_1}k(\cdot,x) - \mathbb{E}_{x\sim\mu_2}k(\cdot,x)\|_{\mathcal{H}} \le \sqrt{C}\|\mu_1-\mu_2\|_{TV}.$$

The derivative F'_μ depends only on f and θ. Thus, $L_{\mu\mu} = L_{\mu h} = 0$. Further, we have

$$\|-f_1(\cdot) + l(\theta_1;\cdot) - (-f_2(\cdot) + l(\theta_2;\cdot))\|_{L^\infty} \tag{105}$$

$$\le \sqrt{C}\|f_1-f_2\|_{\mathcal{H}} + L_0\|\theta_1-\theta_2\|_{\mathcal{H}}, \tag{106}$$

i.e., $L_{f\mu} = \sqrt{C}, L_{f\theta} = L_0$. Here we used that

$$\|f_2(\cdot) - f_1(\cdot)\|_{L^\infty} = \sup_x |f_2(x) - f_1(x)| = \sup_x \langle f_2 - f_1, \phi(x)\rangle_{\mathcal{H}}$$

$$\le \|f_2-f_1\|_{\mathcal{H}} \cdot \sup_x \|\phi(x)\|_{\mathcal{H}} \le \sqrt{C}\cdot\|f_2-f_1\|_{\mathcal{H}}. \tag{107}$$

Finally, the derivative F'_h depends only on f. Thus, $L_{h\theta} = L_{h\mu} = L_{hh} = 0$. Further, we have

$$\|-\epsilon f_1(\cdot) - (-\epsilon f_2(\cdot))\|_{\mathcal{H}} \le \epsilon\|f_1-f_2\|_{\mathcal{H}}. \tag{108}$$

Thus, $L_{hf} = \epsilon$.

As we see, or main assumptions in Theorem B.5 hold for the DRO problem (10). Moreover, stochastic derivatives in (99) and (100) satisfy Assumption B.6. Indeed, we have

$$\mathbb{E}_{X\sim\mu}\|\mathbb{E}_{x\sim\mu}\nabla_\theta l(\theta;x) - \nabla_\theta l(\theta;X)\|_2^2 \le \mathbb{E}_{X\sim\mu}\|\nabla_\theta l(\theta;X)\|_2^2 \le L_0^2, \tag{109}$$

$$\mathbb{E}_{\hat{X}\sim\hat{\mu},X\sim\mu}\left\|\epsilon h(\cdot) + k(\cdot,\hat{X}) + k(\cdot,X) - (\mathbb{E}_{x\sim\hat{\mu}}k(\cdot,x) + \epsilon h(\cdot) - \mathbb{E}_{x\sim\mu}k(\cdot,x))\right\|_{\mathcal{H}}^2 \tag{110}$$

$$\le 2\mathbb{E}_{X\sim\mu}\|k(\cdot,X)\|_{\mathcal{H}}^2 + 2\mathbb{E}_{\hat{X}\sim\hat{\mu}}\left\|k(\cdot,\hat{X})\right\|_{\mathcal{H}}^2 \le 4C. \tag{111}$$

This allows us to apply also Theorem B.7 to the DRO problem (10). This proves Corollary 5.1.