Deep Residual Learning via Large Sample Mean-Field Stochastic Optimization∗

Lijun Bo † Agostino Capponi ‡ Huafu Liao §

Abstract

We study a class of stochastic optimization problems of the mean-field type arising in the optimal training of a deep residual neural network. We estimate the training weights of the network as the optimal relaxed control of a sampling problem, where a population risk criterion is minimized. We establish the existence of optimal relaxed controls when the training set has finite size. The core of our paper is to prove, via Γ-convergence, that the minimizer of the sampled relaxed problem converges to that of the limiting optimization problem, as the number of training samples grows large. We connect the limit of the sampled objective functional to the unique solution, in the trajectory sense, of a nonlinear Fokker-Planck-Kolmogorov (FPK) equation in a random environment.

AMS 2000 subject classifications: 93E35, 93E20, 60F05

Keywords and phrases: Deep residual neural networks; mean-field stochastic optimization; relaxed controls; Fokker-Planck-Kolmogorov equation; Gamma-convergence.

1 Introduction

We study the training performance of a deep residual neural network (RNN), as the number of training samples grows large. RNNs were first introduced in the influential study of He et al. (2016). Their crucial observation was that deep convolutional neural networks are exposed to a degradation problem. As the network depth increases, the training accuracy degrades rapidly. Interestingly, such degradation is not due to overfitting, and adding more layers to a deep model may lead to a higher training error. To ease the training of deep networks and prevent the gradient vanishing problem, He et al. (2016) enhance feedforward neural networks with “shortcut connections”, where such connections implement an identity mapping and their outputs are subsequently added to the outputs of stacked layers.

Our work leverages the optimal control problem formulation for the training weights of a deep RNN, first proposed in E (2017). He uses a continuous time deterministic dynamical system to model the nonlinear transformations across hidden layers of the deep RNN. E (2017) considers an idealized network, which consists of a continuum of layers. The objective of learning the network weights from a finite training set is cast as a sampling control problem, where a population risk criterion is minimized. The control parameters, i.e., the network weights, depend on the population distribution of input/output target pairs. E et al. (2018) provide a rigorous analysis of the control problem formulated in E (2017).

In our paper, we extend their setup significantly to account for regularization during the training process. This yields stochastic dynamics and mean-field interactions for the network state. Our study contributes both to the modeling of residual neural networks (RNNs) and to the performance analysis of such networks when the number of training samples is large. From a modeling perspective, our innovation is to incorporate two common types of regularization into the learning process. The first type is regularization by noise, and consists in injecting noise into the transition function at each layer of the network. This limits the amount

∗We are grateful to Nicholas Garcia Trillos and Weinan E for valuable feedback.
†Email: lijunbo@ustc.edu.cn, School of Mathematical Sciences, University of Science and Technology of China, Hefei, Anhui Province, 230026, China
‡E-mail: ac3827@columbia.edu, Department of Industrial Engineering and Operations Research, Columbia University, New York, 10027, NY, USA.
§E-mail: mathuaf@nus.edu.sg, Department of Mathematics, National University of Singapore, Singapore 119076, Singapore.
of information carried by the units, and acts as a form of regularization because it averages over local neighborhoods of the transition function. We refer to Dieng et al. (2018) for a noise injection function which is fully consistent with ours, and for the quantification of such a regularization effect.\footnote{There exist other forms of regularization by noise that have been considered in the literature. Those include dropout (e.g. Srivastava et al. (2014) and Ba and Frey (2013)), where only a random subset of units is selected in each layer of the network. Other studies regularize the network by adding noise to the weights of a neural network (e.g. Wan et al. (2013) and Kingma et al. (2015)).} The second type of regularization captured by our framework is referred to as \textit{batch normalization} (Ioffe and Szegedy (2015)). This technique was designed to remedy against the fact that the distribution of each layer’s inputs changes during training, as the parameters of previous layers change. Batch normalization mitigates this problem, and accelerates convergence during training of a deep neural network. This is achieved via a normalization step that fixes the means and variances of layer inputs. Such normalization requires sampling inputs from a distribution, and leads to mean-field interactions in the dynamics of the deep RNN.\footnote{For state-of-the-art image classification, batch normalization has been shown to achieve the same accuracy using 14 times fewer training steps than in the model without it.} As a result, the dynamics of the neural network depends not only on the current state, but also on the distribution law of the state.

Once those two regularization techniques are accounted for, the learning process of the deep RNN is described by a system of Brownian driven stochastic differential equations (SDEs) of the mean-field type. We formulate the problem of estimating the trainable weights of the network in terms of finding the optimal relaxed control process of a sampled stochastic control problem. We study the relation between a stochastic residual network with finite training size, and the corresponding network when the number of training samples grows to infinity. Our main result is to show that, as the training size goes to infinity, the sequence of relaxed controls which minimize the finite-dimensional sampled objective functional converges to the relaxed control minimizer of the limiting objective functional. To the best of our knowledge, this convergence result is novel in the literature, and has not been established even in the case where the underlying state dynamics are purely diffusive and have no mean-field interaction.

We next elucidate on the main technical steps leading to our result. First, we prove the existence of optimal relaxed solutions to the finite-dimensional sampled optimization problem stated in Eq. (15) by establishing the precompactness of the minimizing sequence associated with (15). For any relaxed control $Q_N$ of a network with finite sample size $N$, we use the coordinate process consisting of Brownian drivers, input-output training samples and controllable weight process $\theta_N$ to generate a sequence of empirical measure-valued processes $\mu^N_{\text{emp}}$ (see Eq. (20)). We then show that this process admits a limit, which can be characterized by the joint distribution of the initial sample data, the control process and the unique solution of a nonlinear FPK equation in a random environment. We introduce the limiting objective functional $J$, and define it in terms of the joint law of initial data, control process and network state process. The main result presented in Theorem 5.1 is as follows: as $N \to \infty$, the sequence of relaxed control minimizers of the sampled objective functional $J_N$ converge to the relaxed control minimizer of the limiting objective functional $J$. To prove this result, we show that $J_N \Gamma$-converges to $J$ on the space of relaxed controls.

\textbf{Related Literature.} Our work contributes to the growing literature on optimal control approaches to deep learning, where deep RNNs are idealized as continuous-time dynamical systems. The work by Thorpe and van Gennip (2020) provide a discrete-to-continuum Gamma-convergence result for the objective function of a RNN, albeit in a deterministic setup. They consider a training set of finite size, and show that the objective function of the RNN converges, as the number of layers goes to infinity, to a variational problem specified by a set of nonlinear differential equations. Rather than comparing the discretized multi-layer and continuous layer neural network as in Thorpe and van Gennip (2020), we start with a regularized continuous-layer RNN modeled through a SDE of the mean-field type. We use $\Gamma$-convergence to prove that the optimal control in the finite sample case converge to that of the limiting sample objective functional, as the sample size grows large. Jabir et al. (2019) use a neural ODE to describe the learning process of a network. They derive the Pontryagin’s optimality principle for the relaxed control problem which yields the weights of a network with finite sample size, and then provide convergence guarantees for the gradient descent algorithm. In a recent study, Cuchiero et al. (2020) also view deep neural networks as discretizations of controlled ODEs.
(2018) view single-layer neural networks as interacting particle systems. Therein, to overcome the difficulty of minimizing the training error over the set of parameters, they study the minimum of the loss function over its empirical distribution. A related study by Sirignano and Spiliopoulos (2020) models single-layer networks as limits of the sequence of empirical measures of interacting particles. They derive an evolution equation for the empirical measure, and update parameters through the stochastic gradient descent algorithm (SGD). Chizat and Bach (2018) consider a neural network with a single large hidden layer. They study the performance of a non-convex particle gradient descent algorithm as the number of particles grows large, and assume a quadratic loss function to be minimized. Mei et al. (2018) study a two-layer neural network, and show that the dynamics of stochastic gradient descent algorithms used for parameter learning can be approximated by a nonlinear partial differential equation (PDE).

Other studies have considered deeper neural networks. Du et al. (2019) study a deep overparameterized RNN, and prove that gradient descent achieves zero training loss. Sirignano and Spiliopoulos (2020) characterize multi-layer neural networks as the number of hidden layers grows large, and the number of stochastic gradient descent (SGD) iterations grows to infinity. We refer to Mallat (2016) for a mathematical treatment of deep networks, with a focus on convolutional architectures.

Relaxed controls were first studied in El Karoui et al. (1987). Using Krylov’s Markovian selection theorem, Haussmann and Lepeltier (1990) establish the existence of Markovian feedback policies for relaxed optimal control problems. The relaxed stochastic maximum principle was developed, respectively in singular and partially observed optimal control problems, by Bahlali et al. (2007) and Ahmed and Charalambous (2013). More recently, relaxed controls have been applied to analyze existence of Markovian equilibria and relaxed ε-Nash equilibrium of mean field games, both in the presence of idiosyncratic and common noise; see Lacker (2015), Lacker (2016) and Carmona et al. (2016).

The remainder of this paper is organized as follows. In Section 2, we discuss how the problem of finding the training weights of a deep regularized RNN can be cast as the solution to a relaxed control problem of the mean-field type. In Section 3, we establish the existence of optimal solutions to the sampled relaxed optimization problem. In Section 4, we study the convergence of the sampled objective functional. In Section 5, we show that the sequence of minimizers of sampled objective functionals converge to minimizer of the limiting objective functional. Some proofs and additional auxiliary results are delegated to an Appendix.

2 Training Deep RNN via Stochastic Optimization

2.1 Dynamical System Representation of Deep RNN

This section recalls the standard architecture of a deep residual network, and illustrates how its idealized version with infinite depth can be modeled using a continuous time dynamical system.

We begin by considering a simplified version of \( n \)-layered ResNet architecture as in Thorpe and van Gennip (2020). For \( l = 0, 1, \ldots, n-1 \), let \( d \) be the number of neurons in each layer, and \( X^{(n)}(l) \in \mathbb{R}^d \) the states of neurons in layer \( l \). We use \( w^{(n)}(l) \in \mathbb{R}^{d \times d} \) to denote the matrix whose weights determine how neurons in layer \( l \) activate neurons in layer \( l+1 \), and \( b^{(n)}(l) \in \mathbb{R}^d \) to denote the bias vector at layer \( l \). Then, the feed-forward propagation in the \( n \)-layer ResNet model can be represented by the difference equation:

\[
X^{(n)}(l+1) = X^{(n)}(l) + \frac{1}{n} \sigma \left( l, w^{(n)}(l)X^{(n)}(l) + b^{(n)}(l) \right), \quad l = 0, 1, \ldots, n-1, \tag{1}
\]

where \( \frac{1}{n} \) is the scaling factor, and \( \sigma(l, x) = (\sigma_1(l, x_1), \ldots, \sigma_d(l, x_d))^\top \) is the activation function which effectively turns neurons “on” or “off” at layer \( l + 1 \) based on the value of input \( x = (x_1, \ldots, x_d) \) at layer \( l \). In existing implementations, the activation function is chosen to be smooth approximation of a rectified linear activation function or a sigmoid function. Eq. (1) highlights the residual property of the network: the endogenous input at layer \( l+1 \) consists of the non-transformed input \( X^{(n)}(l) \) from layer \( l \), plus a nonlinear transformation of \( X^{(n)}(l) \). The term \( X^{(n)}(l) \) represents information from the previous layer “skipping the processing associated with the layer \( l' \), and being transmitted to layer \( l+1 \) without being transformed. Let \( \theta^{(n)}(l) := (w^{(n)}(l), b^{(n)}(l)) \in \mathbb{R}^{d \times d} \times \mathbb{R}^d \), and write

\[
f(l, \theta^{(n)}(l), x) := \sigma \left( l, w^{(n)}(l)x + b^{(n)}(l) \right). \tag{2}
\]
As observed in prior studies (e.g. E (2017) and Lu et al. (2018)), one can turn the setting described above into an explicit Euler characterization of the ODE, yielding

\[ dX^\theta(t) = f(t, \theta(t), X^\theta(t))dt, \quad t \in (0, T], \]

where \( X^\theta(t) \in \mathbb{R}^d \), and the time step \( dt \) corresponds to the scaling factor \( \frac{1}{T} \). The initial state \( X^\theta(0) = X(0) \) collects the exogenous training inputs. Above, we have used the superscript \( \theta \) to emphasize the dependence of output neurons on the weight matrix and, for a finite \( T \), \( X^\theta \) and \( \theta \) denote real valued functions on \([0, T]\). The argument described above has been used to justify neural network architectures arising from discretization of differential equations (e.g. Chen et al. (2018) and Thorpe and van Gennip (2020)). In the rest of this paper, we will treat (3) as a differential equation, i.e., assume an infinitesimal time step. This effectively corresponds to the limit of an infinitely deep residual network, i.e., consisting of infinitely many layers.

### 2.2 Regularized Deep RNN

We enhance the dynamical system representation of a deep RNN sketched in the previous section to incorporate the most prominent regularization techniques in the training process. Specifically, we allow for (i) additive noise injection into the deterministic hidden units to prevent overfitting, and (ii) batch normalization in each layer to accelerate convergence of the training algorithm. This turns the continuous time dynamics given by (3) into a stochastic dynamical system of the mean-field type.

Concretely, we mimic a supervised learning process, and draw \( N \) i.i.d. training samples

\[ \zeta^i := (X^i(0), Y^i(0)) \in \Xi_K, \]

where \( \Xi_K := [-K, K]^{2d} \) for a globally positive constant \( K \). This means that the training samples are assumed to have compact support. The \( i \)-th training sample consists of the training input \( X^i(0) \), and the label part \( Y^i(0) \). The specific distributional assumptions and their implications are discussed in Section 4.2. We then specify the feed-forward propagation dynamics originated from each training input as follows. Given an original probability space \((\Omega, \mathcal{F}, \mathbb{P}, \mathbb{P})\) with filtration \( \mathbb{F} = (\mathcal{F}_t)_{t \in [0, T]} \) satisfying the usual conditions, the feed-forward network originated from the \( i \)-th training input is the \( d \)-dimensional state process \( X^{\theta,i} = (X^{\theta,i}(t))_{t \in [0, T]} \) whose dynamics is

\[ dX^{\theta,i}(t) = f \left( t, \theta(t), X^{\theta,i}(t), \frac{1}{N} \sum_{j=1}^{N} \rho(X^{\theta,j}(t)) \right) dt + \varepsilon^i dW^i(t), \quad i = 1, \ldots, N, \]

and the initial condition is \( X^{\theta,i}(0) = X^i(0) \). In (5), \( \rho : \mathbb{R}^d \rightarrow \mathbb{R}^q \) is the (vector) batch function, and \( \varepsilon^i \in \mathbb{R}^{d \times p} \). For \( i = 1, \ldots, N \), \( W^i = (W^i(t))_{t \in [0, T]} \) are independent \( p \)-dimensional Brownian motions. The \( \mathbb{F} \)-adapted process \( \theta = (\theta(t))_{t \in [0, T]} \) is the control strategy which will be optimally chosen to minimize the objective criterion introduced later in this section.

The time index \( t \in [0, T] \) represents the depth of the network. The mapping \( f : [0, T] \times \mathbb{R}^m \times \mathbb{R}^d \times \mathbb{R}^q \rightarrow \mathbb{R}^d \) is the vector of activation functions (see, e.g. Haykin (2009) and He et al. (2016)). We denote by \( X^{\theta,i}(t) \) the state at layer \( t \) of the feed-forward transformation triggered by the \( i \)-th training input \( X^i(0) \). The endogenous output of the network is \( X^{\theta,i}(T) \). One can think of an arbitrary layer as receiving samples from a distribution that is determined by the layer below. Such a distribution changes during the course of training, making any layer except the first one responsible not only for learning a good representation, but also for adapting to a changing input distribution.

We next illustrate how the proposed framework can be used to describe continuous layer idealizations of residual deep neural networks at different levels of complexity.

- The function \( f \) depends only on the input variables \((t, \theta, x)\), and \( \varepsilon^i = 0 \). This yields a feedforward propagation in a deep network, where the randomness only comes from the initial input \( X^i(0) \). Such a dynamical system specification recovers the framework of E et al. (2018).

- The function \( f \) depends on all its inputs \((t, \theta, x, \eta)\), and \( \varepsilon^i = 0 \). We next illustrate how, in addition to the features above, this specification incorporates batch normalization (see also Algorithm 1 in
Ioffe and Szegedy (2015)) through the function $\rho: \mathbb{R}^d \to \mathbb{R}^2$. For the purpose of illustration, we set $d = 1$. The batch normalization transform applied to the activation $X^{\theta,i}(t)$ over a batch of size $N$ is given by

$$
\gamma \frac{X^{\theta,i}(t) - \frac{1}{N} \sum_{j=1}^{N} X^{\theta,j}(t)}{\sqrt{\frac{1}{N} \sum_{j=1}^{N} (X^{\theta,i}(t))^2 - \left(\frac{1}{N} \sum_{j=1}^{N} X^{\theta,j}(t)\right)^2 + \iota}} + v,
$$

where $\gamma, v \in \mathbb{R}$ are, respectively, the scale and shift factors and $\iota > 0$ is a constant added to the batch variance to avoid numerical instability. Take the (vector) batch function $\rho(x) = (\rho_1(x), \rho_2(x)) = (x, x^2)$. Then, we may rewrite the above batch normalization as:

$$
\gamma \frac{X^{\theta,i}(t) - \frac{1}{N} \sum_{j=1}^{N} \rho_1(X^{\theta,j}(t))}{\sqrt{\frac{1}{N} \sum_{j=1}^{N} \rho_2(X^{\theta,j}(t)) - \left(\frac{1}{N} \sum_{j=1}^{N} \rho_1(X^{\theta,j}(t))\right)^2 + \iota}} + v.
$$

Because of such normalization, the state dynamics of the neural network become of the mean-field type. For $(t, \theta, x, \eta) = (t, (w, b), x, (\eta_1, \eta_2)) \in [0, T] \times \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}^2$, this yields

$$
f(t, (w, b), x, (\eta_1, \eta_2)) = S\left(t, \frac{\gamma(x - \eta_1)}{\sqrt{\eta_2 - \eta_1^2}} + wv + b\right),
$$

where $x \to S(t, x) = \frac{1}{1 + e^{-x}}$ is the logistic activation function at layer $t$.

- The function $f$ depends on all its inputs $(t, \theta, x, \eta)$, and $\varepsilon^i \neq 0$. This is the most general setting, in which a noise process $W = (W^1(t), \ldots, W^N(t))_{t \in [0, T]}$ is added along the trajectory of deep neural network. The noise injection function is given by the addition of noise to the output in each layer, after applying the transformation of inputs from the layer below. We sample noise from a zero-mean Gaussian distribution so that the noise transition function is unbiased. This means that, conditional on the input $X^{\theta,i}(t)$, the output $X^{\theta,i}(t + dt)$ at layer $t + dt$ coincides on average with the output of a deterministic residual neural network. In other words, the noise-injected transition function preserves, on average, the transition function of the underlying deterministic network. These types of unbiased regularization have been shown to perform favorably compared to classical regularization techniques such as dropout, in the context of recurrent neural networks (see Dieng et al. (2018)).

We impose the following assumptions to ensure that the sampled controlled system described by (5) is well posed:

\begin{enumerate}[(i)]
  \item there exists a global constant $K > 0$ such that $|\varepsilon^i| \leq K$ for all $i \in \mathbb{N}$;
  \item the function $[0, T] \ni t \to f(t, 0, 0, 0)$ is square integrable;
  \item the function $f(t, \theta, x, \eta) : [0, T] \times \mathbb{R}^m \times \mathbb{R}^d \times \mathbb{R}^q$ is Lipschitz continuous in $(\theta, x, \eta)$ uniformly in $t$, i.e., for $(\theta_1, x, \eta), (\theta_2, y, \xi) \in \mathbb{R}^m \times \mathbb{R}^d \times \mathbb{R}^q$,

  $$
  |f(t, \theta_1, x, \eta) - f(t, \theta_2, y, \xi)| \leq |\theta_1|_{\text{Lip}}|\theta_1 - \theta_2| + |x - y| + |\eta - \xi|,
  $$

  where $|f|_{\text{Lip}}$ is the Lipschitz coefficient of $f$ which is independent of the layer index $t$;
  \item the batch function $\rho: \mathbb{R}^d \to \mathbb{R}^q$ is Lipschitz continuous.
\end{enumerate}

The Lipschitz conditions imposed in (iii)-(iv) of the assumption $(A_{\varepsilon,f,\rho})$ guarantee the existence of a unique strong solution of the stochastic system (5). Observe that the batch function $\rho_2(x) = |x|^2$ introduced in Algorithm 1 of Ioffe and Szegedy (2015) is only locally Lipschitz. Nevertheless, for practical matters, we can well approximate introduce $\rho_2$ with the truncation function $\rho_2(M)(x) := x^2 1_{|x| \leq M} + M^2 1_{|x| > M}$ where $M > 0$ is sufficiently large. We can then consider an approximating system to (5), in which $\rho_2$ is replaced by $\rho_2(M)$. Such a system is well-posed. Denote by $X^{\theta,i,M} = (X^{\theta,i,M}(t))_{t \in [0, T]}$ the solution of this approximating system. First, in terms of (8), for any $\theta \in \Theta$ (a compact subset of $\mathbb{R}^m$) and $x^i = (x^i_1, \ldots, x^i_d) \in \mathbb{R}^d (i = 2015)$, the batch function $\rho_2(M)(x) := x^2 1_{|x| \leq M} + M^2 1_{|x| > M}$ where $M > 0$ is sufficiently large.
1, \ldots, N), we have that \( a_i(t, (x^1, \ldots, x^N)) := f(t, \theta, x^i, (\frac{1}{N} \sum_{j=1}^{N} \rho_1(x^j), \frac{1}{N} \sum_{j=1}^{N} \rho_2(x^j))) \) satisfies the linear growth condition uniformly in \( t \in [0, T] \). Since the locally Lipschitz continuity implies strong uniqueness, we have \( X^{\theta, i}(t) = X^{\theta, i, M}(t) \) for \( t \in [0, T] \) a.s., where \( \tau_{M} := \inf\{ t \geq 0 ; |X^{\theta, i}(t)| \geq M \} \). Note that \( \lim_{M \to +\infty} \tau_{M} = +\infty \), a.s.. Therefore \( |X^{\theta, i, M}(t) - X^{\theta, i}(t)| \to 0 \), as \( M \to \infty \), a.s.. Hence, \( \lim_{M \to +\infty} X^{\theta, i, M}(t) \) for \( t \in [0, T] \) is the unique solution of (5).

2.3 Sampled Control Problem

We show how the problem of learning the parameters of the idealized residual neural network can be cast as a dynamic optimization problem. Using the residual neural network dynamics given in (5), we first formulate the strict sampled optimization problem, and then introduce the corresponding relaxed version.

The parameter process \( \theta = (\theta(t))_{t \in [0, T]} \) takes values on a set \( \Theta \subset \mathbb{R}^m \) in an admissible set \( \mathbb{U}^{p,F} \), and is obtained by minimizing the following population risk minimization criterion:

\[
\inf_{\theta \in \mathbb{U}^{p,F}} J_N(\theta) := \inf_{\theta \in \mathbb{U}^{p,F}} \mathbb{E} \left[ L_N(\hat{X}(T), \hat{Y}(0)) + \int_0^T R_N(\theta(t), \theta'(t); \hat{X}(t), \hat{Y}(0)) dt \right],
\]

where \( \hat{X}(t) := (X^{\theta, 1}(t), \ldots, X^{\theta, N}(t)) \) is the feed-forward propagation at \( t \) originated from the training input \( X(0) := (X^1(0), \ldots, X^N(0)) \), \( \hat{Y}(0) := (Y^1(0), \ldots, Y^N(0)) \), and \( \theta'(t) \) denotes the first-order weak derivative w.r.t \( t \) if \( \theta \in \mathcal{H}^1_{m} \). Here, we use \( \mathcal{H}^1_{m} \) to represent the Sobolev space \( W^{2,p}((0, T); \mathbb{R}^m) \) for \( p \in \mathbb{N} \). For \( (\hat{x}, \hat{y}, \theta, \theta') \in (\mathbb{R}^d)^N \times (\mathbb{R}^d)^N \times \mathbb{R}^m \times \mathbb{R}^m \),

\[
L_N(\hat{x}, \hat{y}) := \frac{1}{N} \sum_{i=1}^{N} L(x^i, y^i), \quad R_N(\theta(t), \theta'(t); \hat{x}, \hat{y}) := \frac{1}{N} \sum_{i=1}^{N} R(\theta(t), \theta'(t); x^i, y^i).
\]

Above, \( L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+ \) denotes the terminal loss function, and \( R : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+ \) is the regularizer of the control problem. Observe that, differently from E et al. (2018), the probability law of the state process \( (X^{\theta, 1}, \ldots, X^{\theta, N}) \) enters explicitly into the state forward dynamics.

We assume \( (L, R) \) to be of the squared form, which is often the case in practice (see, e.g. Hasan and Roy-Chowdhury (2015)). For concreteness, given \( \alpha, \beta, \lambda_1, \lambda_2 > 0 \), we define

\[
L(x, y) := \alpha |x - y|^2, \quad R(\theta(t), \theta'(t); x, y) := \lambda_1 |\theta|^2 + \lambda_2 |\theta'|^2 + \beta |x - y|^2,
\]

for \( (x, y, \theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^m \). If \( \lambda_1 = \lambda_2 \), then the regularizer of the control problem (10) includes a \( \mathcal{H}^1_{m} \)-regularizer. Taking the \( \mathcal{H}^1_{m} \)-regularizer into account, the admissible set \( \mathbb{U}^{p,F} \) is defined as:

\[
\mathbb{U}^{p,F} := \left\{ \theta \in L^2(\Omega; \mathcal{H}^1_{m}); \theta \text{ is } F\text{-adapted and } \theta \in \Theta, \text{ a.s. on } (0, T) \times \Omega \right\}.
\]

The admissible set \( \mathbb{U}^{p,F} \) is usually referred to as the strict control set under the original probability space. Throughout the paper, we assume that the state space of the parameter process \( \theta \) satisfies the following property:

(\( \mathbf{A}_\Theta \)) \( \Theta \subset \mathbb{R}^m \) is a compact set, which is \textit{not} necessarily convex.

The objective of our paper is to establish convergence of the minimizer of the sampled objective functional \( J_N \) to the minimizer of a limiting objective functional \( J \). There are two main challenges. First, under (\( \mathbf{A}_\Theta \)), the problem (10) is a control problem with a non-convex policy space, hence in general a global optimum \( \theta^* \in \mathbb{U}^{p,F} \) is not guaranteed to exist. Specifically, the existence of an optimal solution to the optimization problem (10) does not follow from standard compactness techniques used in deterministic optimization problems. Second, even if the set \( \Theta \) were convex, the convergence of the sequence of strict control minimizers to the strict control minimizer of the limiting optimization problem is not guaranteed. This is because the related minimizing sequence of strict controls in general is \textit{not} precompact in \( L^2(\Omega; \mathcal{H}^1_{m}) \).
2.4 Sampled Relaxed Control Problem

We use relaxed controls to bypass the key technical challenges described at the end of the previous section. The set of relaxed controls is the space of probability measures on training samples \( \zeta, \) weight process \( \theta \) and Brownian noise \( W \). This is always a convex set, and compact with weak topology under our standing assumption that \( \Theta \) is compact. As a result, a global solution to the relaxed control problem is always guaranteed to exist. Moreover, the minimizing sequence of our relaxed control problem is precompact in the space of probability measures under weak topology. We are then able to establish convergence of the sequence of (relaxed) minimizers of the sampled objective functional \( J_N \) to the minimizer of the limiting objective functional \( J \), when \( N \) is high. We will provide an expression for the limiting objective functional \( J \) (see Eq. \( (64) \)), and relate its representation to the solution of a FPK equation in a random environment.

We begin by establishing a canonical measurable space \( (\Omega_\infty, \mathcal{F}_\infty) \) via an infinite product space such that the coordinate process \((\zeta, W, \theta) = ((\zeta_i)_{i=1}^\infty, (W_i)_{i=1}^\infty, \theta)\) can be formulated under \((\Omega_\infty, \mathcal{F}_\infty)\). We consider the complete natural filtration \( \mathcal{F} = (\mathcal{F}^\zeta_iW_i, \mathcal{F}^\theta_i)_{t \in [0, T]} \) generated by \((\zeta, W, \theta)\), which is the completion of the filtration flow \( \sigma(\zeta) \vee \sigma(W(s), \theta(s)); \ s \leq t \) for \( t \in [0, T] \). Then, the set \( Q(\nu) \) of relaxed controls is a collection of probability measures \( Q \) on \((\Omega_\infty, \mathcal{F}_\infty)\) such that, under \((Q, \mathcal{F})\), the initial sample data \( \zeta \) has a given law \( \nu \); \( W \) is a sequence of Wiener processes; and \( \theta \in \mathbb{U}^{p, \mathcal{F}} \). We will formally define \( Q(\nu) \) in Definition 3.1. Changing the control variable \( \theta \in \mathbb{U}^{p, \mathcal{F}} \) in \((10)\) to \( Q \in Q(\nu) \), the sampled objective functional is accordingly given by

\[
J_N(Q) := \mathbb{E}^{Q} \left[ L_N(X^\theta(T), \tilde{Y}(0)) + \int_0^T R_N(\theta(t), \theta'(t); \tilde{X}^\theta(t), \tilde{Y}(0)) dt \right],
\]

where, for \( i = 1, \ldots, N \), the state process \((X^\theta, i(t))_{t \in [0, T]} \) is the strong solution to Eq. \((5)\) driven by \((\zeta, W, \theta)\). Then, the relaxed control problem in the finite sample case can be formulated as

\[
\alpha_N := \inf_{Q \in Q(\nu)} J_N(Q).
\]

We call \( Q^* \in Q(\nu) \) an optimal (relaxed) solution of the optimization problem \((15)\) if \( J_N(Q^*) = \alpha_N \).

**Remark 2.1.** For a fixed sample size \( N \), we only use the first \( N \) components of the vector \((\zeta, W) = (\zeta^i, W^i)_{i=1}^\infty \) in the sampled objective functional \( J_N(Q) \), where \( Q \in Q(\nu) \). As \( N \) tends to infinity, the sampled objective functional \( J_N(Q) \) will eventually include all components of \((\zeta, W) = (\zeta^i, W^i)_{i=1}^\infty \).

3 Optimal Relaxed Solutions of Sampled Optimization Problem

This section establishes the existence of optimal relaxed solutions for the sampled optimization problem \((15)\). We work with the canonical space representation, and define the set of relaxed controls \( Q(\nu) \).

The canonical probability space needs to be established in terms of an infinite product space. To wit, define

\[
\Omega_\infty := \mathbb{E}_K^N \times \mathcal{C}_p^1 \times \mathcal{H}^1_m, \quad \mathcal{F}_\infty := \mathcal{B}(\Omega_\infty),
\]

where \( \mathcal{C}_d := C([0, T]; \mathbb{R}^d) \) denotes the space of continuous functions from \([0, T]\) to \( \mathbb{R}^d \) equipped with the uniform norm \( \|h\|_T := \sup_{t \in [0, T]} |h(t)| \) for any \( h \in \mathcal{C}_d \). For \( \zeta = (\zeta^i)_{i=1}^\infty \) and \( W = (W^i)_{i=1}^\infty \), we use \((\zeta, W, \theta)\) to denote the identity map on \( \Omega_\infty \). Let \( \mathcal{F} = (\mathcal{F}^\zeta_iW_iW, \mathcal{F}^\theta_i)_{t \in [0, T]} \) be the complete natural filtration generated by \((\zeta, W, \theta)\). It follows from the Sobolev embedding theorem (see, e.g. Evans (2010)) that \( h \in \mathcal{H}^m_\infty \) if and only if \( h \) equals (dt-a.e.) to an absolutely continuous function whose ordinary derivative (which exists dt-a.e.) belongs to \( \mathcal{E}_m^p \). Here, for \( p \geq 1 \), \( \mathcal{E}_m^p := L^p([0, T]; \mathbb{R}^m) \) equipped with the \( L^p \)-norm \( \|h\|_{\mathcal{E}_m^p} := \left\{ \int_0^T |h(t)|^p \, dt \right\}^{1/p} \) for any \( h \in \mathcal{E}_m^p \). Then, we treat \( \mathcal{H}^m_\infty \) as a subset of \( \mathcal{C}_m \). We next endow the space \( \Omega_\infty \) with the following metric: for \((\gamma, w, \theta) \) and \((\gamma, \tilde{w}, \tilde{\theta}) \in \Omega_\infty \), define

\[
d_m((\gamma, w, \theta), (\gamma, \tilde{w}, \tilde{\theta})) := d_1(\gamma, \gamma) + d_2(w, \tilde{w}) + d_3(\theta, \tilde{\theta}).
\]
4.3

Limit of Large Sampled Optimization Problem

This section studies the convergence of the sampled objective functional \( J_N \) given by (14) as the sample size \( N \) tends to infinity. We first establish a general convergence result for a class of empirical processes which arise in our sampled controlled dynamic model. The convergence result is then used to (i) establish the limiting behavior of \( J_N \) in Section 4.3; and (ii) prove the Gamma-convergence of \( J_N \) to \( J \) in Section 5.

4.1 Convergence of Empirical Processes for Large Samples

We analyze the convergence properties for a class of empirical processes arising in our sampled controlled dynamic model (5), as the sample size \( N \) approaches infinity.

For \( N \in \mathbb{N} \), let \( Q_N \in \mathcal{Q}(\nu) \) where \( \nu \in \mathcal{P}(\mathbb{R}^2) \) is the initial sample law. Let \( (\xi_N, \hat{W}_N, \theta_N) \) be the coordinate process corresponding to \( Q_N \) as in Definition 3.1. Moreover, let \( X_N = (X_N^1(t), \ldots, X_N^n(t))_{t \in [0,T]} \) be a solution of the following SDE:

\[
dX_N^i(t) = f \left( t, \theta_N(t), X_N(t), \frac{1}{N} \sum_{j=1}^N \rho(X_N^j(t)) \right) dt + \varepsilon dW_N^i(t). \tag{19}
\]

In other words, \( X_N^i \) satisfies the SDE (5) driven by \((\xi_N, \hat{W}_N, \theta_N)\). Define \( E := \mathbb{R}^{d \times p} \times \mathbb{R}^d \times \mathbb{R}^d \) and introduce the following empirical measure-valued process given by

\[
\mu_N(t) := \frac{1}{N} \sum_{i=1}^N \delta_{(\xi_N^i, X_N^i(t))}, \quad t \in [0,T]. \tag{20}
\]

Above, for \( i \geq 1 \), \( \xi_N^i := (\varepsilon^i, Y_N^i(0)) \in \mathbb{R}^{d \times p} \times \mathbb{R}^d \). We will show later that, for \( N \geq 1 \), we can view \( \mu_N = (\mu_N(t))_{t \in [0,T]} \) as a sequence of \( \hat{S} := C([0,T]; \mathcal{P}_2(E)) \)-valued random variables, where \( \mathcal{P}_2(E) \) is the
p-order Wasserstein space with underlying metric space \((E, d_E)\). We also recall that \(C([0, T]; \mathcal{P}_2(E))\) is the space of continuous \(\mathcal{P}_2(E)\)-valued functions defined on \([0, T]\). For \(N \geq 1\), we define the following joint distribution:

\[
Q^N := Q_N \circ (\mu^N(0), \theta_N, \mu^N)^{-1}.
\]

(21)

The main result of this section is to characterize the limiting behavior of the sequence of joint laws \((Q^N)^N_{N=1}\) (see Theorem 4.1). This is recovered as the unique solution of a FPK equation with random environment. To start with, let \(D := C_0^\infty(\mathbb{R}^m)\) be the space of test functions with its dual space given by \(D'\). We introduce a related parameterized operator defined on \(D\). Formally, for \((\theta, \eta) \in C_m \times \mathbb{R}^q\), and \((s, e) = (s, (\varepsilon, y, x)) \in [0, T] \times E\), define

\[
\mathcal{A}^{\theta, \eta}_s \varphi(s, e) := \nabla_x \varphi(s, e) + f(s, \theta(s), x, \eta) \nabla_x \varphi(s, e) + \frac{1}{2} \text{tr} \left[ \varepsilon \varepsilon^\top \nabla^2_{xx} \varphi(s, e) \right], \quad \varphi \in D.
\]

(22)

Above, \(\nabla_x \varphi := \left( \frac{\partial \varphi}{\partial x_1}, \ldots, \frac{\partial \varphi}{\partial x_d} \right)^\top\) and \(\nabla^2_{xx} \varphi := \left( \frac{\partial^2 \varphi}{\partial x_i \partial x_j} \right)_{i=1,\ldots,d}^{r=1,\ldots,q}\).

**Theorem 4.1.** Let \((A_{\varepsilon, f, \rho})\) and \((A_{\Theta})\) hold. Suppose further that, for some \(\theta_0 \in \mathcal{P}(\mathcal{P}_2(E) \times C_m)\),

\[
Q_N \circ (\mu^N(0), \theta) N^{-1} \Rightarrow \theta_0, \quad N \to \infty.
\]

(23)

Then \((Q^N)^N_{N=1}\) defined by (21) converges in \(\mathcal{P}(\mathcal{P}_2(E) \times C_m \times \hat{S})\). Moreover, if the law of a \(\mathcal{P}_2(E) \times C_m \times \hat{S}\)-valued r.v. \((\hat{\mu}_0, \hat{\theta}, \hat{\mu})\) defined on some probability space \((\Omega, \hat{\mathcal{F}}, \hat{\mathbb{P}})\) is a limit point of \((Q^N)^N_{N=1}\), then, \(\hat{\mathbb{P}}\)-a.s., \(\hat{\mu}\) is the unique solution to the following FPK equation in a random environment: \(\hat{\mu}(0) = \hat{\mu}_0\), and for \(t \in [0, T]\),

\[
\langle \hat{\mu}(t), \varphi(t) \rangle - \langle \hat{\mu}(0), \varphi(0) \rangle - \int_0^t \left( \hat{\mu}(s), \mathcal{A}^{\hat{\theta}(s), \hat{\mu}(s), \rho}_s \varphi(s) \right) ds = 0, \quad \forall \varphi \in D.
\]

(24)

The roadmap of the proof of Theorem 4.1 consists of three steps as follows:

(i) We prove the precompactness of the marginal distributions \((Q^N)^N_{N=1}\) in \(\mathcal{P}_2(\hat{S})\), where

\[
Q^N_\mu := Q_N \circ (\mu^N)^{-1};
\]

(25)

It thus follows from (23) that \((Q^N)^N_{N=1}\) is tight.

(ii) We then prove that, for any weak limit point of a convergent subsequence of \((Q^N)^N_{N=1}\) of the form

\[
\hat{\mathbb{P}} \circ (\hat{\mu}_0, \hat{\theta}, \hat{\mu})^{-1}, \quad \hat{\mu}\text{ is the unique solution of a FPK equation in a random environment with initial condition } \hat{\mu}_0, \hat{\mathbb{P}}\text{-a.s.}
\]

(iii) Finally, we show that \((Q^N)^N_{N=1}\) admits a unique weak limit point.

We next give a lemma (whose proof is reported in the Appendix), which will be used to verify the relative compactness of \((Q^N)^N_{N=1}\) in \(\mathcal{P}(\hat{S})\) needed to prove step (i).

**Lemma 4.2.** Let \((A_{\varepsilon, f, \rho})\) and \((A_{\Theta})\) hold. Let \(\varepsilon > 0\). Then, it holds that

\[
\lim_{M \to \infty} \sup_{N \geq 1} Q^N_\mu \left( \left\{ \theta \in \hat{S} : \sup_{t \in [0, T]} \int_E |\varepsilon|^{2+\varepsilon} \theta(t, dc) \geq M \right\} \right) = 0,
\]

(26)

and for any \(\varepsilon > 0\),

\[
\lim_{\delta \to 0} \sup_{N \geq 1} Q^N_\mu \left( \left\{ \theta \in \hat{S} : \sup_{|t-s| \leq \delta} \mathcal{W}_{E, 2}(\theta(t), \theta(s)) > \varepsilon \right\} \right) = 0.
\]

(27)

Above, \(\mathcal{W}_{E, 2}\) is the quadratic Wasserstein metric on \(\mathcal{P}_2(E)\).

The following lemma completes step (i).
Lemma 4.3. Let \((\mathbf{A}_{\varepsilon,f,\rho})\) and \((\mathbf{A}_0)\) hold. Then, the sequence of marginal distributions \((Q_N^\mu)_{N=1}^\infty\) defined by (25) is relatively compact in \(\mathcal{P}_2(\hat{S})\).

Proof. We first verify that \((Q_N^\mu)_{N=1}^\infty \subset \mathcal{P}_2(\hat{S})\). Let \(e = (\varepsilon, y, x)\) and \(\hat{e} = (\hat{\varepsilon}, \hat{y}, \hat{x}) \in E\). We take a measure-valued process \(\hat{\vartheta} \in \hat{S}\) satisfying \(\sup_{t \in [0,T]} \int_E |\hat{e}|^2 \hat{\vartheta}(t, d\hat{e}) < +\infty\). We endow \(\hat{S}\) with the metric

\[
d_{\hat{S}}(\hat{\vartheta}, \hat{\vartheta'}) := \sup_{t \in [0,T]} \mathcal{W}_{E,2}(\vartheta(t), \hat{\vartheta}(t)), \quad \hat{\vartheta}, \hat{\vartheta'} \in \hat{S}.
\]

(28)

Then, for any \(N \geq 1\), we have that

\[
\int_{\hat{S}} d_{\hat{S}}^2(\hat{\vartheta}, \hat{\vartheta})Q_N^\mu(d\hat{\vartheta}) = \mathbb{E}^Q \left[ d_N^2(\mu^N, \hat{\vartheta}) \right] = \mathbb{E}^Q \left[ \sup_{t \in [0,T]} \mathcal{W}_{E,2}^2(\mu^N(t), \hat{\vartheta}(t)) \right]
\]

\[
\leq \mathbb{E}^Q \sup_{t \in [0,T]} \left[ \int_{E \times E} |\vartheta - \hat{\vartheta}|^2 \mu^N(t, de) \hat{\vartheta}(t, d\hat{e}) \right] \leq 2\mathbb{E}^Q \sup_{t \in [0,T]} \left[ \int_E |\vartheta|^2 \mu^N(t, de) \right] + 2 \sup_{t \in [0,T]} \int_E |\hat{e}|^2 \hat{\vartheta}(t, d\hat{e}).
\]

(29)

(30)

Since \(\hat{\mu} \in \hat{S}\), the 2nd term on the r.h.s. of the inequality (29) is finite. For the 1st term on the r.h.s. of the inequality (29), using (25) it follows that

\[
\mathbb{E}^Q \left[ \sup_{t \in [0,T]} \int_{E} |\vartheta|^2 \mu^N(t, de) \right] \leq \frac{1}{N} \sum_{i=1}^N \mathbb{E}^Q \left[ \left| \xi^i_N, X^i_N(t) \right|^2 \right]
\]

\[
\leq \frac{1}{N} \sum_{i=1}^N \mathbb{E}^Q \left[ |\xi^i_N|^2 \right] + \mathbb{E}^Q \left[ \sup_{t \in [0,T]} \left| \hat{X}_N(t) \right|^2 \right].
\]

(30)

where we recall that \(\hat{X}_N(t) := (X^1_N(t), \ldots, X^N_N(t))\) for \(t \in [0,T]\). Using \((\mathbf{A}_{\varepsilon,f,\rho}), (\mathbf{A}_0)\) and noting that \(\xi^i \in \mathcal{X}_K\) for all \(i \geq 1\), it follows from (29) that

\[
\sup_{N \geq 1} \int_{\hat{S}} d_{\hat{S}}^2(\hat{\vartheta}, \hat{\vartheta})Q_N^\mu(d\hat{\vartheta}) \leq \frac{2}{N} \sum_{i=1}^N \mathbb{E}^Q \left[ |\xi^i_N|^2 \right] + 2 \sup_{N \geq 1} \mathbb{E}^Q \left[ \sup_{t \in [0,T]} \left| \hat{X}_N(t) \right|^2 \right]
\]

\[
+ 2 \sup_{t \in [0,T]} \int_{E} |\hat{e}|^2 \hat{\vartheta}(t, d\hat{e})
\]

\[
< +\infty.
\]

(31)

This shows that \(Q_N^\mu \in \mathcal{P}_2(\hat{S})\) for all \(N \geq 1\).

We next prove that \((Q_N^\mu)_{N=1}^\infty \subset \mathcal{P}_2(\hat{S})\) is relatively compact. By Theorem 7.12 in Villani (2003), \((Q_N^\mu)_{N=1}^\infty\) is relatively compact in \(\mathcal{P}_2(\hat{S})\) if and only if (I) \((Q_N^\mu)_{N=1}^\infty\) is relative compact in \(\mathcal{P}(\hat{S})\); and (II) \((Q_N^\mu)_{N=1}^\infty\) satisfies the uniform integrability condition, i.e., for some \(\hat{\vartheta} \in \hat{S}\),

\[
\lim_{R \to \infty} \sup_{N \geq 1} \int_{\{\hat{\vartheta} \in \hat{S}; d_{\hat{S}}(\hat{\vartheta}) \geq R\}} d_{\hat{S}}^2(\hat{\vartheta}, \hat{\vartheta})Q_N^\mu(d\hat{\vartheta}) = 0.
\]

(32)

- The proof of (I): \((Q_N^\mu)_{N=1}^\infty\) is relatively compact in \(\mathcal{P}(\hat{S})\).

By Ascoli’s theorem, a subset \(C \subset \hat{S} = C([0,T]; \mathcal{P}_2(E))\) is relatively compact if (I1) for each \(t \in [0,T]\), \(\{\vartheta(t); \vartheta \in C\} \subset \mathcal{P}_2(E)\) is relatively compact; and (I2): \(C\) is equicontinuous under \(W_2\). Moreover, using again Theorem 7.12 in Villani (2003), (I1) holds if and only if (I11) holds: for each \(t \in [0,T]\), \(\{\vartheta(t); \vartheta \in C\}\) is relatively compact in \(\mathcal{P}(E)\); and (I12) holds: the following uniform integrability condition is satisfied:

\[
\lim_{R \to \infty} \sup_{\vartheta \in C} \int_{|e| \geq R} |e|^2 \vartheta(t, de) = 0.
\]

(33)
Let $\epsilon > 0$. For $M, \delta, \varepsilon > 0$, define the following subsets of $\hat{S}$:

$$C_1(M) := \left\{ \theta \in \hat{S} : \sup_{t \in [0, T]} \int_E |e|^{2+\varepsilon} \partial(t, de) \leq M \right\}, \quad C_2(\delta, \varepsilon) := \left\{ \theta \in \hat{S} : \sup_{|t-s| \leq \delta} W_{E, \Delta}(\partial(t), \partial(s)) \leq \varepsilon \right\}. \quad (34)$$

Then, for any $\theta \in C_1(M)$ and $t \in [0, T]$, it follows that

$$\partial(t, B_R(0)) \leq \frac{1}{R^{2+\varepsilon}} \sup_{t \in [0, T]} \int_E |e|^{2+\varepsilon} \partial(t, de) \leq \frac{M}{R^{2+\varepsilon}} \to 0, \quad R \to \infty, \quad (35)$$

where $B_R(0) := \{ e \in E; |e| \leq R \}$ for $R > 0$. On the other hand, it is easy to see that

$$\lim_{R \to \infty} \sup_{\theta \in C_1(M)} \int_{\{e \in E; |e| \geq R\}} |e|^{2+\varepsilon} \partial(t, de) \leq \lim_{R \to \infty} \frac{1}{R^{2+\varepsilon}} \sup_{\theta \in C_1(M)} \int_E |e|^{2+\varepsilon} \partial(t, de) \leq \lim_{R \to \infty} \frac{M}{R^{2+\varepsilon}} = 0. \quad (36)$$

By (I$_{11}$) and (I$_{12}$), this implies that $C_1(M) \subset \hat{S}$ satisfies (I$_1$). To continue, fix $\varepsilon > 0$, by (26) of Lemma 4.2 there exists $N_0 = N_0(\varepsilon) \geq 1$ and $M_0 = M_0(\varepsilon) \geq 1$ such that $\sup_{N \geq N_0} Q_n^\mu(C_1(M)) \leq \frac{\varepsilon}{2}$. Note that $\lim_{M \to \infty} Q_n^\mu(C_1(M)) = 0$ for any $N \geq 1$ by (26). It follows that there exists $M_1 = M_1(\varepsilon)$ large enough such that $\sup_{1 \leq N \leq N_0} Q_n^\mu(C_1(M)) \leq \frac{\varepsilon}{2M_1}$. Then, let $M := M_0 \vee M_1$, and hence $\sup_{N \geq 1} Q_n^\mu(C_1(M)) \leq \varepsilon$. By applying the limiting result (27) of Lemma 4.2, $\lim_{\varepsilon \to 0} \sup_{N \geq 1} Q_n^\mu(C_2(\delta, n^{-1})) = 0$ for each $n \geq 1$. Then, there exists $\delta_n > 0$ satisfying $\lim_{\varepsilon \to 0} \delta_n = 0$ such that $\sup_{N \geq 1} Q_n^\mu(C_2(\delta, n^{-1})) \leq \frac{\varepsilon}{2}$. Define $\mathcal{C} := C_1(M) \cap (\bigcap_{n \geq 1} C_2(\delta_n, n^{-1})) \subset \hat{S}$. Then $\mathcal{C}$ is relatively compact in $\hat{S}$, and it follows from the above given estimates that $\sup_{N \geq 1} Q_n^\mu(\mathcal{C}) \leq 2\varepsilon$. This shows that $(Q_n^\mu)_{n=1}^{\infty}$ is relatively compact in $\mathcal{P}(\hat{S})$.

- The proof of (II): $(Q_n^\mu)_{n=1}^{\infty}$ satisfies the uniform integrability (32).

First of all, for any $N \geq 1$, it holds that

$$\int d_S^{2+\varepsilon}(\theta, \hat{\theta}) Q_n^\mu(d\theta) \leq \mathbb{E}^{Q_n} \left[ \sup_{t \in [0, T]} \left( \int_E |e|^{2+\varepsilon} \mu^N(t, de) \hat{\theta}(t, de) \right)^{\frac{2+\varepsilon}{2}} \right] \leq 2^{1+\varepsilon} \mathbb{E}^{Q_n} \left[ \sup_{t \in [0, T]} \left( \int_E |e|^{2+\varepsilon} \mu^N(t, de) \right)^{\frac{2+\varepsilon}{2}} \right] + 2^{1+\varepsilon} \left( \sup_{t \in [0, T]} \int_E |e|^{2+\varepsilon} \hat{\theta}(t, de) \right)^{\frac{2+\varepsilon}{2}}. \quad (37)$$

It follows from Jensen’s inequality that, for some constant $C_\varepsilon > 0$ which only depends on $\varepsilon$,

$$\sup_{N \geq 1} \mathbb{E}^{Q_n} \left[ \sup_{t \in [0, T]} \left( \int_E |e|^{2+\varepsilon} \mu^N(t, de) \right)^{\frac{2+\varepsilon}{2}} \right] \leq \sup_{N \geq 1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}^{Q_n} \left[ \left( \xi_i^\varepsilon(t), X_N^\varepsilon(t) \right)^{2+\varepsilon} \right] \leq C_\varepsilon \left( \sup_{N \geq 1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}^{Q_n} \left[ |\xi_i^\varepsilon|^{2+\varepsilon} \right] \right). \quad (38)$$

Observe that $\hat{\theta} \in \hat{S}$. Using the assumption (A$_{\varepsilon, f, \rho}$), Lemma A.1 and noting that $\zeta^i \in \Xi_K$ for all $i \geq 1$, it follows from (29) and (37) that, as $R \to \infty$,

$$\sup_{N \geq 1} \int_{\hat{S}} d_S^{2+\varepsilon}(\theta, \hat{\theta}) Q_n^\mu(d\theta) \leq \frac{1}{R^{2+\varepsilon}} \sup_{N \geq 1} \int_{\hat{S}} d_S^{2+\varepsilon}(\theta, \hat{\theta}) Q_n^\mu(d\theta) \to 0, \quad (39)$$

i.e., the uniform integrability (32) holds. This completes the proof of the lemma. \qed

The following proposition completes step (ii).

**Proposition 4.4.** Let the assumptions of Theorem 4.1 hold. Then $(Q_n^\mu)_{n=1}^{\infty} \subset \mathcal{P}(\mathcal{P}_2(E) \times \mathcal{C}_m \times \hat{S})$ is tight. If the law of a $\mathcal{P}_2(E) \times \mathcal{C}_m \times \hat{S}$-valued r.v. $(\hat{\mu}, \hat{\theta}, \hat{\mu})$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is the weak limit of a convergent subsequence of $(Q_n^\mu)_{n=1}^{\infty}$, then, $\mathbb{P}$-a.s. $\hat{\mu}$ is the unique solution of FPK equation (24) with initial condition $\hat{\mu}(0) = \hat{\mu}_0$. 

11
Proof. The tightness of \((Q^N)_{N=1}^\infty\) follows from the condition (23) and Lemma 4.3. Since \(\hat{P} \circ (\mu_0, \hat{\theta}, \hat{\mu})^{-1}\) is the weak limit of a convergent subsequence of \((Q^N)_{N=1}^\infty\), Skorokhod representation theorem implies the existence of a probability space \((\Omega^*, \mathcal{F}^*, \hat{P}^*)\), a sequence of \(\mathcal{P}_2(E) \times \mathcal{C}_m \times \mathcal{S}\)-valued r.v.s \((\mu_0^{N,*}; \theta_N^{N,*}; \mu_0^N)\) and \((\mu_0^{N,*}; \theta^*; \mu_0^N)\) satisfying \(\hat{P} \circ (\mu_0^{N,*}; \theta_N^{N,*}; \mu_0^N)^{-1} = Q^N\); \(\hat{P} \circ (\mu_0^{N,*}; \theta^*; \mu_0^N)^{-1} = \hat{P} \circ (\hat{\mu}(0), \hat{\theta}; \hat{\mu})^{-1}\), and \(\hat{P}^\ast\)-a.s., as \(N \to \infty\),

\[
\mu_0^{N,*} \Rightarrow \mu_0^* \quad \text{in} \quad \mathcal{P}_2(E); \quad \theta_N^{N,*} \to \theta^* \quad \text{in} \quad \mathcal{C}_m; \quad \mu_0^{N,*} \to \mu_0^* \quad \text{in} \quad (\hat{S}, d_\hat{S}). \tag{40}
\]

Moreover, by Itô’ formula and BDG inequality, for any \(p \geq 1\),

\[
\lim_{N \to \infty} \mathbb{E}^\ast \left[ \sup_{t \in [0,T]} \left| \left\langle \mu_0^{N,*}(t), \varphi(t) \right\rangle - \left\langle \mu_0^N(0), \varphi(0) \right\rangle - \int_0^t \left\langle \mu_0^{N,*}(s), \mathcal{A}\mathcal{B}_t^\ast(\mu_0^{N,*}(s), \rho) \varphi(s) \right\rangle ds \right|^{2p} \right] = 0. \tag{41}
\]

We next claim that for any \(t \in [0,T]\) and test function \(\varphi \in D\), \(\hat{P}^\ast\)-a.s.

\[
\lim_{N \to \infty} \mathcal{Y}_{t,\varphi}(\mu_0^{N,*}; \theta_N^{N,*}; \langle \mu_0^{N,*}, \rho \rangle) = \mathcal{Y}_{t,\varphi}(\mu^*, \theta^*; \langle \mu^*, \rho \rangle), \tag{42}
\]

where the mapping \(\mathcal{Y}_{t,\varphi} : \hat{S} \times \mathcal{C}_m \times \mathcal{C}_1 \to \mathbb{R}\) is defined as:

\[
\mathcal{Y}_{t,\varphi}(\mu, \theta, h) := \int_0^t \left\langle \mu(s), \mathcal{A}^{\theta,h}(s) \varphi(s) \right\rangle ds. \tag{43}
\]

Here the definition of \(\mathcal{A}^{\theta,h}(s)\) is given in (22). Then, for any \((\theta_i, h_i) \in \mathcal{C}_m \times \mathcal{C}_1\) with \(i = 1, 2\) and \(\mu \in \hat{S}\), it follows from \((A_{\varepsilon,f,\rho})\) that there exists a constant \(C_\varphi > 0\) such that

\[
|\mathcal{Y}_{t,\varphi}(\mu, \theta_1, h_1) - \mathcal{Y}_{t,\varphi}(\mu, \theta_2, h_2)| \leq C_\varphi \left\| \theta_1 - \theta_2 \right\|_T + \|h_1 - h_2\|_T. \tag{44}
\]

Moreover, by \((A_{\varepsilon,f,\rho})\), \(\mathcal{A}^{\theta,h}(s)\) is \(C_\varphi[1 + \|\theta\|_T + \|h\|_T + |\varepsilon|^2]\). Hence, for any \(s \in [0,T]\), by (40) and Theorem 7.12 of Villani (2003), we arrive at the conclusion that, \(\hat{P}^\ast\)-a.s., \((\mu_0^{N,*}(s), \mathcal{A}^{\theta,h}(s)) \to \langle \mu^*(s), \mathcal{A}^{\theta,h}(s) \rangle\) as \(N \to \infty\). Further, for \(q(\varepsilon) := |\varepsilon|^2\), it holds that, \(\hat{P}^\ast\)-a.s.

\[
\sup_{N \geq 1} \left| \left\langle \mu_0^{N,*}(s), \mathcal{A}^{\theta,h}(s) \varphi(s) \right\rangle \right| \leq C_{\varphi,h} + C_{\varphi,h} \sup_{N \geq 1} \left\langle \mu_0^{N,*}(s), q \right\rangle = C_{\varphi,h} + C_{\varphi,h} \sup_{N \geq 1} W_2(\mu_0^{N,*}(s), \delta_0)
\]

\[
\leq C_{\varphi,h} \sup_n W_2(\mu^*(s), \delta_0) + C_{\varphi,h} \sup_{N \geq 1} d_{\mathcal{S}}(\mu_0^{N,*}, \mu^*), \tag{45}
\]

for some positive constant \(C_{\varphi,h}\) which is independent of \(N\). Using the limiting results given in (40), we then obtain that \(d_{\mathcal{S}}(\mu_0^{N,*}, \mu^*) \to 0\) as \(N \to \infty\), \(\hat{P}^\ast\)-a.s. Note that \(\mu^* \in \hat{S}\). Then, it follows from the dominated convergence theorem that, for \(t \in [0,T]\),

\[
\lim_{N \to \infty} \int_0^t \left\langle \mu_0^{N,*}(s), \mathcal{A}^{\theta,h}(s) \varphi(s) \right\rangle ds = \int_0^t \left\langle \mu^*(s), \mathcal{A}^{\theta,h}(s) \varphi(s) \right\rangle ds, \quad \hat{P}^\ast\text{-a.s.} \tag{46}
\]

Using (44), (40) and (46), we deduce that, \(\hat{P}^\ast\)-a.s.

\[
\left| \mathcal{Y}_{t,\varphi}(\mu_0^{N,*}; \theta_N^{N,*}; \langle \mu_0^{N,*}, \rho \rangle) - \mathcal{Y}_{t,\varphi}(\mu^*, \theta^*; \langle \mu^*, \rho \rangle) \right| \leq |\mathcal{Y}_{t,\varphi}(\mu_0^{N,*}; \theta_N^{N,*}; \langle \mu_0^{N,*}, \rho \rangle) - \mathcal{Y}_{t,\varphi}(\mu_0^{N,*}; \theta^*; \langle \mu^*, \rho \rangle)|
\]

\[
+ |\mathcal{Y}_{t,\varphi}(\mu_0^{N,*}; \theta^*; \langle \mu^* \rho \rangle) - \mathcal{Y}_{t,\varphi}(\mu^*, \theta^*; \langle \mu^*, \rho \rangle)|
\]

\[
\to 0, \quad N \to \infty. \tag{47}
\]

This proves the limit in (42). By applying Fatou’s lemma, (41) and (42), we obtain that

\[
\hat{E} \left[ \sup_{t \in [0,T]} \left| \left\langle \hat{\mu}(t), \varphi(t) \right\rangle - \left\langle \hat{\mu}(0), \varphi(0) \right\rangle - \int_0^t \left\langle \hat{\mu}(s), \mathcal{A}^{\theta,h}(\hat{\mu}(s), \rho) \varphi(s) \right\rangle ds \right|^{2p} \right] = 0.
\]
\[ E^* \sup_{t \in [0,T]} \left[ \left\{ \mu^*(t), \varphi(t) \right\} - \left\{ \mu^*(0), \varphi(0) \right\} - \int_0^t \left\{ \mu^*(s), A^{\varphi^*,(\mu^*(s),\rho)} \varphi(s) \right\} ds \right]^{2p} \]

\[ \leq \liminf_{N \to \infty} E^* \sup_{t \in [0,T]} \left[ \left\{ \mu^{N^*}(t), \varphi(t) \right\} - \left\{ \mu^{N^*}(0), \varphi(0) \right\} - \int_0^t \left\{ \mu^{N^*}(s), A^{\varphi^*,(\mu^{N^*}(s),\rho)} \varphi(s) \right\} ds \right]^{2p} = 0. \]

This proves (24) for all \( \omega \in \Omega_0 \) with some \( \Omega_0 \subset \Omega \) satisfying \( \hat{F}(\Omega_0) = 1 \).

We next prove the uniqueness of a solution to the FPK equation (24) in the trajectory sense. This can be done by verifying the conditions (DH1)-(DH4) imposed in Theorem 4.4 of Manita et al. (2015). To this purpose, for fixed \( \omega \in \Omega_0 \), and for \( (t, e, \mu) \in [0,T] \times E \times \mathcal{P}_2(E) \), define

\[
\begin{aligned}
A(e) &:= \frac{1}{2} \left( \begin{array}{c}
0 \\
\varepsilon
\end{array} \right), \quad b_\omega(t, e, \mu) := \left( f(t, \theta^*(t, \omega), x, (\mu, \rho)), 0 \right) \\
L_\omega(\mu) &:= \text{tr}[A(e) \nabla^2 e] + b_\omega(t, e, \mu)^T \nabla e.
\end{aligned}
\]

It then follows from (49) that \( \sqrt{A(e)} \) is twice differentiable in \( e \), and hence the assumption (DH1) in Theorem 4.4 of Manita et al. (2015) is satisfied. Choose the convex function \( \Phi \in C^2(E) \) as \( \Phi(e) = 1 + |e|^2 \) for \( e \in E \). For any \( \dot{e} = (\dot{e}, \dot{e}, \dot{x}, \dot{x}) \in E \), it follows from (A\( e,f,\varphi \)) that there exists a constant \( C_{f,\varphi} > 0 \) (which may vary from line to line) such that

\[
(b_\omega(t, e + \dot{e}, \mu) - b_\omega(t, e, \mu))^T \dot{e} = (f(t, \theta^*(t, \omega), x + \dot{x}, (\mu, \rho)) - f(t, \theta^*(t, \omega), x, (\mu, \rho)))^T \dot{x}
\]

\[
\leq C_{f,\varphi} |\dot{e}|^2 \leq C_{f,\varphi} \Phi(\dot{e}),
\]

and there exists a constant \( C_{f,\varphi,\mu,\omega} > 0 \) such that

\[
L_\omega(\mu) \Phi(e) = \text{tr}[A(e) \nabla^2 e \Phi(e)] + b_\omega(t, e, \mu)^T \nabla e \Phi(e) \leq C_{f,\varphi,\mu,\omega} \Phi(e).
\]

It also follows from (A\( e,f,\varphi \)) that

\[
\frac{|f(t, \theta^*(t, \omega), x, (\mu, \rho))|^2}{\Phi(e)} \leq 1 + \frac{||\theta^*(\omega)||^2}{\Phi(e)} + \frac{|\mu|}{\Phi(e)} + \frac{1}{\Phi^2(e)} \leq 1 + \frac{||\theta^*(\omega)||^2}{\Phi} + + |(\mu, \rho)|^2.
\]

This verifies the condition (DH2). For any \( \mu, \nu \in \mathcal{P}_2(E) \), it follows from (49) and (A\( e,f,\varphi \)) that

\[
|b_\omega(t, e, \mu) - b_\omega(t, e, \nu)| = |f(t, \theta^*(t, \omega), x, (\mu, \rho)) - f(t, \theta^*(t, \omega), x, (\nu, \rho))|
\]

\[
\leq |f|_{\text{Lip}} |(\mu, \rho) - (\nu, \rho)| = |f|_{\text{Lip}} |\rho|_{\text{Lip}} \left| \mu - \nu, \rho \right|_{\text{Lip}} \leq |f|_{\text{Lip}} |\rho|_{\text{Lip}} |W_{E,2}(\mu, \nu) =: G(W_{E,2}(\mu, \nu)),
\]

where \( G(y) := |f|_{\text{Lip}} |\rho|_{\text{Lip}} y \) for \( y \in [0, \infty) \). Obviously, the function \( G \) is continuous and increasing on \([0, \infty)\) with \( G(0) = 0 \). This verifies the condition (DH3). Next, we verify the condition (DH4). Take \( \Psi(e) = \Phi(e) \) for \( e \in E \). Hence, \( \Psi \in C^2(E) \), \( \Psi \geq 1 \), and \( |\nabla \Psi(e)| + |\nabla^2 \Psi(e)| \leq C \) for some constant \( C > 0 \). Moreover, it holds that \( \lim_{|e| \to \infty} \Psi(e) = +\infty \). We deduce from (49) that, for any \( \mu \in \mathcal{P}_2(E) \),

\[
\frac{|A(e) \nabla \Psi(e)|}{\Psi(e)} + \frac{\sqrt{A(e) \nabla \Psi(e)}}{\Psi^2(e)} + \frac{|L_\omega(\mu) \Psi(e)|}{\Psi(e)} \leq C_{f,\varphi,\mu,\omega} \Phi(e).
\]

For fixed \( \omega \in \Omega_0 \), the uniqueness of a solution to (24) follows from Theorem 4.4 in Manita et al. (2015).

The next lemma concludes the step (iii) outlined in the proof roadmap. Its proof follows directly from the Gluing lemma (see, e.g. Lemma 7.6 in Villani (2003)) and is omitted here.

**Lemma 4.5.** Let assumptions of Theorem 4.1 hold. Then, the precompact sequence \((Q_N^e)_{N=1}^\infty\) has a unique weak limit point.
We now have all the ingredients to prove the main result (Theorem 4.1) of this section.

**Proof of Theorem 4.1.** It follows from Proposition 4.4 and Lemma 4.5 that \((Q^N)_{N=1}^∞ \subset \mathcal{P}(\mathcal{P}_2(E) \times \mathcal{C}_m \times \mathcal{S})\) is convergent under the weak topology. Let us endow \(O := \mathcal{P}_2(E) \times \mathcal{C}_m \times \mathcal{S}\) with the following metric: for \(a_i = (θ_{0i}, w_i, δ_i) \in O\) with \(i = 1, 2\),

\[
d_{O}(a_1, a_2) := W_{E,2}(θ_{01}, θ_{02}) + \|w_1 - w_2\|_T + d_S(δ_1, δ_2).
\]

(55)

Then, using assumptions \((A_{ε,f,μ})\) and \((A_ε)\), the fact that \(ζ_N = (ζ_N^i)_{i=1}^∞ \in \Xi_K^N\), and Lemma A.1, for \(δ = (δ_0, 0, δ_0) \in O\) and \(ε > 0\), it follows that

\[
\sup_{N \geq 1} \int_{\{O: d_{O}(a, δ) \geq R\}} d^2_{O}(a, δ)Q^N(\text{d}a) \leq C \sup_{N \geq 1} \mathbb{E}^{Q^N} \left[ \left| W_{E,2}^2(μ^N(0), δ_0) + \|θ_N\|_T^2 + d^2_S(μ^N, δ_0) \right|^{\frac{2+ε}{2}} \right]
\]

\[
\leq C \sup_{N \geq 1} \mathbb{E}^{Q^N} \left[ \frac{1}{N} \sum_{i=1}^N (|ε_i|^2 + |ζ_N^i|^2) + \|θ_N\|_T^2 + \frac{1}{N} \sum_{i=1}^N \|X_N^i\|_T^2 \right] \leq C_{K,ε,T} < +∞,
\]

(56)

where \(C\) and \(C_{K,ε,T}\) are some positive constants independent of \(N\). This implies that

\[
\lim_{R \to ∞} \sup_{N \geq 1} \int_{\{O: d_{O}(a, δ) \geq R\}} d^2_{O}(a, δ)Q^N(\text{d}a) = 0.
\]

(57)

Then, the convergence of \((Q^N)_{N=1}^∞ \subset \mathcal{P}_2(O)\) follows from Theorem 7.12 in Villani (2003) along with the uniform integrability result given in (57).

Lastly, we show the uniqueness of the weak limit point of the marginal distributions \((Q^N_μ)_{N=1}^∞ \subset \mathcal{P}_2(\mathcal{S})\) defined by (25). Lemma 4.3 shows that \((Q^N_μ)_{N=1}^∞ \subset \mathcal{P}_2(\mathcal{S})\) is precompact. The following corollary is an immediate consequence of Theorem 4.1.

**Corollary 4.6.** Let assumptions of Theorem 4.1 hold. Then, the precompact sequence \((Q^N_μ)_{N=1}^∞ \subset \mathcal{P}_2(\mathcal{S})\) has a unique limit \(Q^*_μ \in \mathcal{P}_2(\mathcal{S})\) satisfying \(W_{S,2}(Q^N_μ, Q^*_μ) \to 0\) as \(N \to ∞\).

### 4.2 A Sufficient Condition for Weak Convergence (23)

This section provides an easily verifiable sufficient condition on the initial sample law \(ν ∈ \mathcal{P}(\Xi_K^N)\) that guarantees the weak convergence (23) assumed in Theorem 4.1:

\((A_ν)\) For \(N ∈ \mathbb{N}\), define the mapping \(I_N : \Xi_K^N \to \mathcal{P}_2(E)\) as follows: for any \(ζ = (ζ^i, Y^i)_{i=1}^∞ \in \Xi_K^N\), \(I_N(ζ) := \frac{1}{N} \sum_{i=1}^N δ_{ζ^i, Y^i, X^i}\). Then, there exists a measurable mapping \(I_* : \Xi_K^N \to \mathcal{P}_2(E)\) such that

\[
ν \left\{ \left\{ \hat{ζ} ∈ \Xi_K^N; \lim_{N \to ∞} W_{E,2}(I_N(\hat{ζ}), I_*(\hat{ζ})) = 0 \right\} \right\} = 1.
\]

(58)

The following remark presents an example of initial laws of training samples that satisfy \((A_ν)\):

**Remark 4.7.** Consider any sequence of i.i.d. \(\Xi_K\)-valued r.v.s \((X^i, \hat{Y}^i)_{i=1}^∞\) on some probability space \((Ω, \mathcal{F}, \mathbb{P})\). Set \(\hat{ζ} = (\hat{ζ}^i)_{i=1}^∞ = (X^i, \hat{Y}^i)_{i=1}^∞\) and hence \(\hat{ζ} ∈ \Xi_K^N\). Then, for any sequence \((ε^i)_{i=1}^∞\) satisfying \(\lim_{i \to ∞} ε^i = ε^*\), the law of large number (LLN) yields \(\overset{\mathbb{P}}{\lim}_{N \to ∞} W_{E,2}(I_N(\hat{ζ}(ω)), I_*(\hat{ζ}(ω))) = 0\). In this specific setup, \(I_* := δ_{(ε^*, γ, σ^*)} \otimes \mathbb{P}O(\hat{ζ}^i)^{-1}\). Consider the initial sample law \(ν := \mathbb{P}O(\hat{ζ}^i)^{-1}\), then it holds that

\[
ν \left\{ \left\{ \hat{ζ} ∈ \Xi_K^N; \lim_{N \to ∞} W_{E,2}(I_N(\hat{ζ}), I_*) = 0 \right\} \right\} = \overset{\mathbb{P}}{\lim}_{N \to ∞} W_{E,2}(I_N(\hat{ζ}(ω)), I_*) = 0\}
\]

(59)

Hence, the assumption \((A_ν)\) is satisfied in this specific setup.

Note that the convergence of relaxed controls \((Q_N)_{N=1}^∞\) does not imply the weak convergence (23). The reason is that the empirical distribution of the initial data may not converge. Condition \((A_ν)\) guarantees that the distribution of initial data is well behaved. The following lemma, proven in the Appendix, shows that if the convergence of \((Q_N)_{N=1}^∞\) converges, then \((A_ν)\) implies the weak convergence in (23).

14
Lemma 4.8. Let \((A_i)\) hold. Consider an arbitrary sequence \((Q_N)_{N=1}^{\infty}, Q \subset Q(\nu)\) with \(\lim_{N \to \infty} W_{\Omega_{\infty},2}(Q_N, Q) = 0\), then \(Q_N \circ (\zeta_N, I_N(\zeta_N), \theta_N)^{-1} \implies Q \circ (\zeta, I(\zeta), \theta)^{-1}\), \(N \to \infty\). Here, \((\zeta, W, \theta)\) (resp. \((\zeta_N, W_N, \theta_N)\)) is the coordinate process corresponding to \(Q\) (resp. \(Q_N\)).

4.3 Convergence of Sampled Objective Functionals

In this section, we prove the convergence, as the number of samples \(N \to \infty\), of the sampled objective functional \(J_N(Q)\) given by (14), for a fixed \(Q \in Q(\nu)\). Such an analysis uses the generalized convergence result given in Theorem 4.1.

For fixed \(Q \in Q(\nu)\), let \((\zeta, W, \theta)\) be the canonical (or coordinate) process corresponding to \(Q\). For \(i \geq 1\), recall that \(X^{\theta,i} = (X^{\theta,i}(t))_{t \in [0,T]}\) solves the SDE (5) driven by \((\zeta, W, \theta)\). Next, we introduce a new empirical measure-valued process given by

\[
\hat{\mu}^N(t) := \frac{1}{N} \sum_{i=1}^{N} \delta_{(\epsilon^i, X^{\theta,i}(t))}, \quad \text{for } t \in [0,T].
\]  

(60)

The empirical process \(\hat{\mu}^N = (\hat{\mu}^N(t))_{t \in [0,T]}\) can be viewed as the counterpart of \(\mu^N\) defined in (20), but driven by \((\zeta, W, \theta)\) instead of the sample \((\zeta_N, W_N, \theta_N)\) from \(Q_N \in Q(\nu)\). We then define the law of \(\hat{\mu}^N\) as:

\[
\hat{Q}^N := Q \circ (\hat{\mu}^N)^{-1}.
\]  

(61)

Using (60) and (61), we may rewrite the sampled objective functional \(J_N(Q)\) in (14) as follows:

\[
J_N(Q) = \mathbb{E}^Q \left[ \langle \hat{\mu}^N(T), L \rangle \right] + \frac{\beta}{\alpha} \mathbb{E}^Q \left[ \int_0^T \langle \hat{\mu}^N(t), L \rangle dt \right] + \mathbb{E}^Q \left[ \int_0^T \{ \lambda_1 |\theta(t)|^2 + \lambda_2 |\theta'(t)|^2 \} dt \right]
\]  

(62)

In the above expression, we recall that \(\langle \mu, f \rangle := \int f d\mu\) for \(\mu \in \mathcal{P}(E)\) and the loss function is defined by \(L(e) = \alpha |x - y|^2\) where \(e = (\epsilon, y, x) \in E\). By applying Lemma 4.8 and Corollary 4.6, we immediately get the following result.

Lemma 4.9. Let \((A_{\varepsilon,f,\rho})\), \((A_\theta)\) and \((A_\varphi)\) hold. Then, the precompact sequence \((\hat{Q}^N)_{N=1}^{\infty} \subset \mathcal{P}_2(\hat{S})\) has a unique limit point \(\hat{Q}^* \in \mathcal{P}_2(\hat{S})\) satisfying \(W_{\hat{S},2}(\hat{Q}^N, \hat{Q}^*) \to 0\) as \(N \to \infty\).

Moreover, the limit point \(\hat{Q}^* \in \mathcal{P}_2(\hat{S})\) can be explicitly characterized, as shown in the following lemma whose proof is in the Appendix.

Lemma 4.10. Let \((A_{\varepsilon,f,\rho})\), \((A_\theta)\) and \((A_\varphi)\) hold. Assume \((Q_N)_{N=1}^{\infty}, Q \subset Q(\nu)\) satisfy \(\lim_{N \to \infty} W_{\Omega_{\infty},2}(Q_N, Q) = 0\). Then \(Q^N\) defined by (21) converges to \(Q \circ (I_*, \theta, \mu_*)^{-1}\) in \(\mathcal{P}_2(\mathcal{P}_2(E) \times C_m \times \hat{S})\), as \(N \to \infty\), where \(I_*\) is given in \((A_\varphi)\), and \(Q\)-a.s., \(\mu_*\) is the unique solution of FPK equation: for all \(\varphi \in \mathcal{D}\),

\[
\begin{aligned}
&\left\{ \langle \mu_*(t), \varphi(t) \rangle - \langle \mu_*(0), \varphi(0) \rangle - \int_0^t \langle \mu_*(s), A^\theta(\mu_*(s), \rho) \varphi(s) \rangle ds = 0, \quad t \in (0,T]; \\
&\mu_*(0) = I_*
\right. \end{aligned}
\]  

(63)

Moreover, it holds that \(\hat{Q}^* = Q \circ (I_*, \theta, \mu_*)^{-1}\).

For a given \(Q \in Q(\nu)\) and the unique limit point \(\hat{Q}^* \in \mathcal{P}_2(\hat{S})\) from Lemma 4.10, we define

\[
J(Q) := \int_\hat{S} \langle \vartheta(T), L \rangle \hat{Q}^*(d\vartheta) + \frac{\beta}{\alpha} \int_0^T \left( \int_\hat{S} \langle \vartheta(t), L \rangle \hat{Q}^*(d\vartheta) \right) dt + \mathbb{E}^Q \left[ \int_0^T \{ \lambda_1 |\theta(t)|^2 + \lambda_2 |\theta'(t)|^2 \} dt \right].
\]  

(64)

By Lemma A.1 in the Appendix, we then have that \(\sup_{N \geq 1} J_N(Q) < +\infty\) for each \(Q \in Q(\nu)\). We are now ready to state the main result of this section:
Proposition 4.11. Let \((A_{\varepsilon,f,\rho}), (A_\Theta)\) and \((A_v)\) hold. Then, for any \(Q \in \mathcal{Q}(\nu)\),
\[
\lim_{N \to \infty} J_N(Q) = J(Q),
\]
where \(J_N(Q)\) and \(J(Q)\) for \(Q \in \mathcal{Q}(\nu)\) are defined by (62) and (64) respectively.

Proof. To prove the proposition, we apply Lemma 4.9 and Theorem 7.12 of Villani (2003). Let \(t \in [0,T]\), and define \(\mathcal{L}(\vartheta) := \langle \vartheta(t), L \rangle = \int_E L(e) \vartheta(t, de)\) for all \(\vartheta \in \hat{S} = C([0,T]; \mathcal{P}_2(E))\). First of all, it follows from (28) that
\[
|\mathcal{L}(\vartheta)| = \int_E L(x,y) \vartheta(t, de) \leq 2 \int_E |e|^2 \vartheta(t, de) \leq 2 \sup_{t \in [0,T]} W_{E,2}^2(\vartheta(t), \delta_0) = 2d_S^2(\vartheta, \delta_0),
\]
where \(e = (\varepsilon,y,x) \in V\). This shows that \(\mathcal{L}\) satisfies the growth condition on \(\hat{S}, d_S\) with \(p = 2\) in Theorem 7.12-(iv) of Villani (2003). Next, assume that \((\vartheta_l)_{l \geq 1} \subset \hat{S}\) satisfy \(\vartheta_l \to \vartheta\) on \((\hat{S}, d_S)\), as \(l \to \infty\). This implies that \(\sup_{t \in [0,T]} W_{E,2}(\vartheta_l(t), \vartheta(t)) \to 0\) as \(l \to \infty\). Using Theorem 7.12 of Villani (2003), it follows that for any continuous function \(\phi\) on \(E\) satisfying the quadratic growth, \(\langle \vartheta_l(t), \phi \rangle \to \langle \vartheta(t), \phi \rangle\) as \(l \to \infty\). Note that \(\phi(e) := |x - y|^2 \leq 2|e|^2\) and hence \(\mathcal{L}(\vartheta_l) \to \mathcal{L}(\vartheta)\) as \(l \to \infty\). Thus, we have shown that \(\mathcal{L}\) is continuous and satisfies the quadratic growth on \((\hat{S}, d_S)\). By Lemma 4.9, \(W_{E,2}(\hat{Q}^N, \hat{Q}^\ast) \to 0\) as \(N \to \infty\).

Again, by Theorem 7.12 of Villani (2003), we conclude that
\[
g_N(t) := \int_{\hat{S}} \mathcal{L}(\vartheta) \hat{Q}^N(d\vartheta) \to \int_{\hat{S}} \mathcal{L}(\vartheta) \hat{Q}^\ast(d\vartheta), \quad N \to \infty.
\]
Moreover, for \(\varepsilon > 0\), using Jensen’s inequality and Lemma A.1, we deduce the existence of a positive constant \(C_{\varepsilon,T}\) which only depends on \(\varepsilon, T\) such that
\[
\sup_{N \geq 1} \int_0^T |g_N(t)|^{1+\varepsilon/2} dt = \sup_{N \geq 1} \int_0^T \left[ \mathbb{E}Q \left[ |\hat{\mu}^N(t), L| \right] \right]^{1+\varepsilon/2} dt \leq \sup_{N \geq 1} \mathbb{E}Q \left[ \frac{1}{N} \sum_{t=0}^T \left| X_{\varepsilon,i}(t) - Y_{\varepsilon,i}(0) \right|^{2+\varepsilon} \right] + K^{2+\varepsilon} < +\infty.
\]

This implies that, as \(R \to \infty\),
\[
\sup_{N \geq 1} \int_{\{t \in [0,T] : |g_N(t)| \geq R\}} |g_N(t)| dt \leq \frac{1}{R^{\varepsilon/2}} \sup_{N \geq 1} \int_{\{t \in [0,T] : |g_N(t)| \geq R\}} |g_N(t)|^{1+\varepsilon/2} dt \to 0.
\]

Then, by Vitali’s convergence theorem together with (67) and (69), it follows that
\[
\int_0^T g_N(t) dt = \int_0^T \left( \int_{\hat{S}} \mathcal{L}(\vartheta) \hat{Q}^N(d\vartheta) \right) dt \to \int_0^T \left( \int_{\hat{S}} \mathcal{L}(\vartheta) \hat{Q}^\ast(d\vartheta) \right) dt, \quad N \to \infty.
\]
The desired convergence then follows from (67) and (70), recalling the expressions of \(J_N(Q)\) and \(J(Q)\) given, respectively, by (62) and (64).

5 Convergence of Minimizers of Sampled Objective Functionals

In this section, we show that the sequence of minimizers of the sampled objective functionals converges to the minimizer of the limiting objective functional, if \(N\) is large enough. To establish this result, a key step is to prove the so-called \(\Gamma\)-convergence of \(J_N\) to \(J\) (see (62) and (64)).

Before introducing the main result of this section, we first metrize the space \(\mathcal{Q}(\nu) \subset \mathcal{P}_2(\Omega_\infty)\) by taking the quadratic Wasserstein distance \(W_{1,\infty,2}\) on \(\mathcal{Q}(\nu)\). The main result of the paper is as follows:
Theorem 5.1. Let $(A_{ε,f,ρ})$, $(A_δ)$ and $(A_ω)$ hold. Then, it holds that
\[
\frac{\inf_{Q \in \mathcal{Q}(ν)} J_N(Q)}{\inf_{Q \in \mathcal{Q}(ν)} J(Q)} \rightarrow N \to \infty,
\]
where the minimum of $J(Q)$ over $Q \in \mathcal{Q}(ν)$ exists. Moreover, if the minimizing sequence $(Q_N)_{N=1}^\infty \subset \mathcal{Q}(ν)$ (up to a subsequence) converges to some $Q^* \in \mathcal{Q}(ν)$ in $W_{1,2}$, then $Q^*$ minimizes $J(Q)$ over $Q \in \mathcal{Q}(ν)$.

Proof. The proof of Theorem 5.1 requires proving (i) the Gamma-convergence from $J_N$ to $J$, which is done in Proposition 5.2; and (ii) that the minimizing sequence of the sampled optimization problem (15) is precompact in $W_{1,2}$, which is shown in Lemma 5.3.

We next give the definition of Gamma-convergence of the sequence of sampled objective functionals $(J_N)_{N=1}^\infty$ on $(\mathcal{Q}(ν), W_{1,2}(N))$ (see, e.g. Dal Maso (1993)):

Definition 5.1. $J_N : \mathcal{Q}(ν) \to \mathbb{R}$ Gamma-converges to some functional $J : \mathcal{Q}(ν) \to \mathbb{R}$, i.e., $J = \Gamma \lim_{N \to \infty} J_N$ on $\mathcal{Q}(ν)$, if the following conditions hold:

(i) \textbf{(liminf inequality):} For any $Q \in \mathcal{Q}(ν)$ and every sequence $(Q_N)_{N=1}^\infty$ converging to $Q$ in $(\mathcal{Q}(ν), W_{1,2}(N))$, we have that $\liminf_{N \to \infty} J_N(Q_N) \geq J(Q)$;

(ii) \textbf{(limsup inequality):} For any $Q \in \mathcal{Q}(ν)$, there exists a sequence $(Q_N)_{N=1}^\infty$ which converges to $Q$ in $(\mathcal{Q}(ν), W_{1,2}(N))$ (this sequence is said to be a $Γ$-realising sequence), such that $\limsup_{N \to \infty} J_N(Q_N) \leq J(Q)$.

The following proposition shows that $J_N$ Gamma-converges to $J$ as $N \to \infty$.

Proposition 5.2. Let $(A_{ε,f,ρ})$, $(A_δ)$ and $(A_ω)$ hold. Then $J = \Gamma \lim_{N \to \infty} J_N$ on $(\mathcal{Q}(ν), W_{1,2}(N))$.

Proof. Let $Q \in \mathcal{Q}(ν)$ and take $Q_N = Q$ for all $N \geq 1$. Then, it follows from Proposition 4.11 that $\lim_{N \to \infty} J_N(Q_N) = J(Q)$. Therefore, $(Q_N)_{N=1}^\infty \subset \mathcal{Q}(ν)$ is a $Γ$-realising sequence. Hence, the limsup inequality in Definition 5.1 holds.

It remains to prove the liminf inequality. Let $(Q_N)_{N=1}^\infty, Q \subset \mathcal{Q}(ν)$ satisfy $\lim_{N \to \infty} W_{1,2}(N,Q) = 0$. Then, it follows from Lemma 4.10 that $Q_N = Q \circ (\mu_N(N), 0, \theta_N, N) \to Q \circ \mu_\ast(N, \mu_\ast) = \mu_\ast$ in $\mathcal{P}_2(\mathcal{P}_2(E) \times C_m \times \hat{S})$. The exact expression of $(I_\ast, \mu_\ast)$ is given in Lemma 4.10. Recall the expression of $Q_N$ given in (25). Then $Q_N$ converges to $\hat{Q}_\ast := Q \circ \mu_\ast(\hat{S})$, as $N \to \infty$. Using similar arguments to those in the proof of Proposition 4.11, this leads to
\[
\lim_{N \to \infty} \int_S (\vartheta(T,L)) Q_N \left( d\vartheta \right) = \lim_{N \to \infty} \frac{β}{α} \left( \int_0^T \left( \int_S \vartheta(t,L) Q_N \left( d\vartheta \right) \right) dt \right).
\]

We next state and prove the following claim:
\[
\ell := \liminf_{N \to \infty} E^{Q,N} \left[ \left( \|\theta_N\|_{L^2_m}^2 + \|\theta_N'\|_{L^2_m}^2 \right) \right] \geq \sup_{\epsilon, f, \rho} E^{Q,N} \left[ \left( \|\theta\|_{L^2_m}^2 + \|\theta'\|_{L^2_m}^2 \right) \right] \geq \ell \leq +\infty.
\]

If $\ell = +\infty$, then (73) trivially holds. If $\ell < +\infty$ (note that $\ell \geq 0$ by the way it is defined), then passing to a subsequence (call it $Q_N$ again), we may assume that
\[
\lim_{N \to \infty} E^{Q,N} \left[ \left( \|\theta_N\|_{L^2_m}^2 + \|\theta_N'\|_{L^2_m}^2 \right) \right] = \ell < +\infty.
\]

Note that $Q_N \circ \mu_\ast^{-1} \rightarrow Q \circ \theta^{-1}$ as $N \to \infty$. Using Skorokhod’s representation theorem, there exists a probability space $(Ω^*, \mathcal{F}^*, \mathbb{P}^*)$, a sequence of $C_m$-valued r.v.s $(\theta_N^*)_{N=1}^\infty$, $\mathbb{P}^*$ such that $\mathbb{P}^* \circ (\theta_N^*)^{-1} = Q \circ \theta^{-1}$, as $N \to \infty$, $\theta_N \to \theta^*$ in $C_m$, $\mathbb{P}^*$-a.s. It follows from the dominated convergence theorem that
\[
\lim_{N \to \infty} E^{Q,N} \left[ \left( \|\theta_N\|_{L^2_m}^2 \right) \right] = \lim_{N \to \infty} E^* \left[ \left( \|\theta_N^*\|_{L^2_m}^2 \right) \right] = E^* \left[ \left( \|\theta^*\|_{L^2_m}^2 \right) \right] = E^{Q,N} \left[ \left( \|\theta\|_{L^2_m}^2 \right) \right].
\]
We next prove that \((\theta_N^*)_{N=1}^{\infty}\) is bounded in \(L^2((0, T) \times \Omega^*; dt \otimes d\nu^*)\). This can be derived using similar arguments to those employed to derive \((A.22)\). For completeness, we provide the precise mathematical details. For any \(\phi \in \mathcal{D}\), define \(\mathcal{T}_N(\phi) := (\theta_N^*, \phi^*)\). Let \((\phi)_{N=1}^{\infty} \subset \mathcal{D}\) be dense in \(L^2_m\). Then, for each \(N \geq 1\),

\[
E^\ast \left[ \sup_{t \geq 1} \frac{\|\theta_N^* \phi_t^*\|_{L^2_m}}{\|\phi_t\|_{L^2_m}} \right] = E^{Q_N} \left[ \sup_{t \geq 1} \frac{\|\theta_N^* \phi_t^*\|_{L^2_m}}{\|\phi_t\|_{L^2_m}} \right] = E^{Q_N} \left[ \|\theta_N^* \|_{L^2_m} \right] < \infty. \tag{76}
\]

By Hahn-Banach theorem and Riesz representation theorem, there exists an \(L^2_m\)-valued r.v. \(\hat{\theta}_N^*\), such that \(\mathcal{T}_N(\phi) = (\hat{\theta}_N^*, \phi)\) for all \(\phi \in L^2_m\), \(\mathbb{P}^\ast\)-a.s. In particular, \(\mathcal{T}_N(\phi) = (\theta_N^*, \phi^*) = (\hat{\theta}_N^*, \phi)\) for any \(\phi \in \mathcal{D}\). This yields that \(\theta_N^* \in H_m^\ast\), \(\mathbb{P}^\ast\)-a.s. It then follows from \((74)\) that

\[
\sup_{N \geq 1} E^\ast \left[ \|\theta_N^* \|_{L^2_m} \right] = \sup_{N \geq 1} E^\ast \left[ \sup_{t \geq 1} \|\theta_N^* \phi_t^*\|_{L^2_m} \right] = \sup_{N \geq 1} E^{Q_N} \left[ \sup_{t \geq 1} \|\theta_N^* \phi_t^*\|_{L^2_m} \right] = \sup_{N \geq 1} E^{Q_N} \left[ \|\theta_N^* \|_{L^2_m} \right] < \infty. \tag{77}
\]

This shows that \((\theta_N^*)_{N=1}^{\infty}\) is bounded in \(L^2((0, T) \times \Omega^*; dt \otimes d\nu^*)\), and hence \(\theta_N^* \) (up to a subsequence) converges weakly to some \(\hat{\theta}^* \in L^2((0, T) \times \Omega^*; dt \otimes d\nu^*)\) as \(N \to \infty\). As shown in Proposition 3.1, for any \(\phi \in \mathcal{D}\) and \(H \in L^\infty(\Omega^*; \mathbb{P}^\ast)\), by the weak convergence property

\[
E^\ast \left[ (\theta^*, \phi^*) H \right] = \lim_{N \to \infty} E^\ast \left[ (\theta_N^*, \phi^*) H \right] = -\lim_{N \to \infty} E^\ast \left[ (\theta_N^*, \phi^*) H \right] = -E^\ast \left[ (\hat{\theta}^*, \phi) H \right]. \tag{78}
\]

Thus, for any \(\phi \in \mathcal{D}\), \((\theta^*, \phi^*) = - (\hat{\theta}^*, \phi), \mathbb{P}^\ast\)-a.s.. The separability of \(\mathcal{D}\) implies that \(\mathbb{P}^\ast\)-a.s., \((\theta^*, \phi^*) = - (\hat{\theta}^*, \phi), \) for all \(\phi \in \mathcal{D}\), i.e., \(\theta^* = \hat{\theta}^*, \mathbb{P}^\ast\)-a.s.. Similarly to the derivation of the estimates \((A.22)\) and \((77)\), we obtain that

\[
\liminf_{N \to \infty} E^{Q_N} \left[ \|\theta_N^* \|_{L^2_m}^2 \right] \leq \liminf_{N \to \infty} E^\ast \left[ \|\theta_N^* \|_{L^2_m}^2 \right] \geq \liminf_{N \to \infty} E^\ast \left[ \|\theta_N^* \|_{L^2_m}^2 \right] \geq \liminf_{N \to \infty} E^{Q_N} \left[ \|\theta_N^* \|_{L^2_m}^2 \right]. \tag{79}
\]

Thus, the proof of \((73)\) follows immediately from \((75)\) and \((79)\). Finally, note that

\[
J_N(Q_N) = \int_S \left( \langle \partial(T), L \rangle Q^N_\mu (d\theta) + \frac{\beta}{\alpha} \int_0^T \left( \int_S \langle \partial(t), L \rangle Q^N_\mu (d\theta) \right) dt + E^{Q_N} \left[ \|\theta_N \|_{L^2_m}^2 + \|\theta_N^* \|_{L^2_m}^2 \right] \right. \tag{80}
\]

Then, the \(\liminf\) inequality in Definition 5.1 follows from \((64)\), \((72)\), and \((73)\). \(\square\)

To conclude the proof of Theorem 5.1, it remains to establish the relatively compactness of the minimizing sequence of the sampled optimization problem \((15)\). As shown in Proposition 3.1 of Section 3, for each \(N \geq 1\) there exists a relaxed solution \(Q_N \in Q(\nu)\) such that \(J_N(Q_N) = \inf_{Q \in Q(\nu)} J_N(Q)\). We will prove below that such a sequence \((Q_N)_{N=1}^{\infty}\) is precompact under the quadratic Wasserstein distance \(W_{\infty, 2}\).

**Lemma 5.3.** Let \((A_{\varepsilon, f, \phi})\) and \((A_\phi)\) hold. Then, the above minimizing sequence \((Q_N)_{N=1}^{\infty} \subset Q(\nu)\) is precompact in \(W_{\infty, 2}\).

**Proof.** Lemma A.1 implies that \(\sup_{N \geq 1} J_N(Q) \leq \ell\) for some \(\ell > 0\). Together with Proposition 3.1, this implies the existence of a constant \(\ell > 0\) and some \(Q \in Q(\nu)\) such that

\[
J_N(Q_N) = \inf_{Q \in Q(\nu)} J_N(Q) \leq \sup_{N \geq 1} J_N(Q) \leq \ell, \quad \forall N \geq 1. \tag{81}
\]

Therefore, it follows that

\[
\sup_{N \geq 1} E^{Q_N} \left[ \|\theta_N \|_{L^2_m}^2 + \|\theta_N^* \|_{L^2_m}^2 \right] \leq \sup_{N \geq 1} J_N(Q_N) \leq \ell. \tag{82}
\]

Let \(X_N = (\zeta_N, W_N, \theta_N)\) be the canonical process corresponding to \(Q_N\). Using similar arguments to those in the proof of Proposition 3.1, \((Q_N \circ \theta_N^{-1})_{N=1}^{\infty}\) is tight on \(C_m\), and moreover \((Q_N \circ X_N^{-1})_{N=1}^{\infty}\) is tight on \(\Omega_\infty\) because \((\Omega_\infty, d)\) is Polish. Then, by Prokhorov’s theorem, there exists a \(Q^* \in \mathcal{P}(\Omega_\infty)\) such that the minimizing sequence \((Q_N)_{N=1}^{\infty}\), up to a subsequence, converges to \(Q^*\) under weak topology. Using Skorokhod
representation theorem, there exists a probability space \((\Omega^*, F^*, P^*)\), \(X_N^* = (\zeta_N, W_N, \theta_N)\) with \(X_N^* \overset{d}{=} X_N\), and \(X^* = (\zeta^*, W^*, \theta^*)\) with \(P^* \circ (X^*)^{-1} = Q^*\) such that, \(P^*\)-a.s., as \(N \to \infty\),
\[
\zeta_N \to \zeta^* \text{ in } E^N; \quad W_N^* \to W^* \text{ in } C^m_p; \quad \theta_N \to \theta^* \text{ in } C_m.
\]
(83)

Further, we obtain that (i): for \(N \geq 1\), \(\theta_N \in H^m_{\gamma} P^*-\text{a.s. and hence } P^* \circ (X_N^*)^{-1} \in \mathcal{P}(\Omega^*_\infty)\); (ii): \(Q^* = P^* \circ (X^*)^{-1} \in \mathcal{P}(\Omega^*_\infty)\), and hence \(Q^* \in Q(\nu)\). To prove that \(Q_N\) converges to \(Q^*\) as \(N \to \infty\) in \(W_{\Omega^*_\infty, 3}\), using Definition 6.8 and Theorem 6.9 in Villani (2009), it suffices to prove that
\[
\lim_{N \to \infty} \int_{\Omega^*_\infty} d^2_{\gamma}\left((\gamma, w, \vartheta), (\hat{\gamma}, \hat{w}, \hat{\vartheta})\right) Q_N(d(\gamma, w, \vartheta)) = \int_{\Omega^*_\infty} d^2_{\gamma}\left((\gamma, w, \vartheta), (\hat{\gamma}, \hat{w}, \hat{\vartheta})\right) Q^*(d(\gamma, w, \vartheta)),
\]
(84)
for some \((\hat{\gamma}, \hat{w}, \hat{\vartheta}) \in \Omega^*_\infty\). First, note that, for all \(N \geq 1\), it holds that
\[
\int_{\Omega^*_\infty} d^2_{\gamma}\left((\gamma, w, \vartheta), (\hat{\gamma}, \hat{w}, \hat{\vartheta})\right) Q_N(d(\gamma, w, \vartheta)) = \mathbb{E}^*[\left[d_1(\zeta_N^*, \hat{\gamma}) + d_2(W_N^*, \hat{w}) + d_3(\theta_N^*, \hat{\vartheta})\right]^2].
\]
(85)

Observing that, for any \(N \geq 1\), \((\zeta_N^*, W_N^*)\) and \((\zeta^*, W^*)\) are identically distributed by (i) and (ii) of Definition 3.1, it follows that \((d_1^2(\zeta_N^*, \hat{\gamma}) + d_2^2(W_N^*, \hat{w}))_{N=1}^{\infty}\) is uniformly integrable. Using (A_\alpha), we deduce that \((d_3(\theta_N^*, \hat{\vartheta}))_{N=1}^{\infty}\) is uniformly integrable. It then follows from (83) and Vitali’s convergence theorem that
\[
\lim_{N \to \infty} \mathbb{E}^*\left[d_1(\zeta_N^*, \hat{\gamma}) + d_2(W_N^*, \hat{w}) + d_3(\theta_N^*, \hat{\vartheta})\right]^2 = \mathbb{E}^*\left[d_1(\zeta^*, \hat{\gamma}) + d_2(W^*, \hat{w}) + d_3(\theta^*, \hat{\vartheta})\right]^2 = \int_{\Omega^*_\infty} d^2_{\gamma}\left((\gamma, w, \vartheta), (\hat{\gamma}, \hat{w}, \hat{\vartheta})\right) Q^*(d(\gamma, w, \vartheta)),
\]
(86)
which yields (84). This completes the proof of the lemma.

\[\Box\]

**A Appendix**

This Appendix provides proofs to some of the propositions and lemmas stated in the main body. Additionally, it provides a miscellaneous of technical results, along with the corresponding proofs, that are used to derive proofs of propositions and theorems stated in the main body of the paper.

*Proof of Proposition 3.1.* Without loss of generality, we assume that \(\alpha_N = \inf_{Q \in Q(\nu)} J_N(Q) < +\infty\). In view of (14), let \((Q_k)_{k=1}^\infty \subset Q(\nu)\) be a minimizing sequence such that
\[
0 \leq \alpha_N \leq J_N(Q_k) \leq \alpha_N + \frac{1}{k}, \quad \forall k \geq 1.
\]
(A.1)

Above, for \(k \geq 1\), the objective functional
\[
J_N(Q_k) = \mathbb{E}^{Q_k}\left[L_N(\hat{X}\kappa(t), \hat{Y}_k(0)) + \int_0^T R_N(\theta_k(t), \theta_k'(t); \hat{X}\kappa(t), \hat{Y}_k(0))dt\right],
\]
(A.2)
where \(X_k := (\zeta_k, W_k, \theta_k) = ((\zeta_k^i)_{i=1}^{\infty}, (W_k^i)_{i=1}^{\infty}, \theta_k)\) is the canonical process. Since \(Q_k \in Q(\nu)\), it follows from Definition 3.1 that (i) \(Q_k \circ \zeta_k^{-1} = \nu\); (ii) \(W_k\) consists of a sequence of independent Wiener processes on \((\Omega^*_\infty, \mathcal{F}, Q_k)\); (iii) \(\theta_k \in U^{Q_k, \kappa}\). Moreover, for \(i = 1, \ldots, N\), the state process \(X_{\kappa}^{\theta_i}\) is the strong solution of (5) driven by \((\zeta_k, W_k, \theta_k)\). Then, it follows from (11), (A.1) and (A.2) that, for all \(k \geq 1\),
\[
(\lambda_1 \wedge \lambda_2) \mathbb{E}^{Q_k}\int_0^T \{\theta_k(t)^2 + |\theta_k'(t)|^2\}dt \leq J_N(Q_k) \leq \alpha_N + \frac{1}{k}.
\]
(A.3)

This implies that
\[
\sup_{k \geq 1} \mathbb{E}^{Q_k}\left[\int_0^T \{\theta_k(t)^2 + |\theta_k'(t)|^2\}dt\right] \leq \frac{\alpha_N}{\lambda_1 \wedge \lambda_2}.
\]
(A.4)
Note that $\theta_k$ is a $\mathcal{H}_{m}^{1}$ (as a subset of $\mathcal{C}_{m}$)-valued random variable on $(\Omega_{\infty}, \mathcal{F}_{\infty}, Q_k)$ for $k \geq 1$. Then, for any $\delta > 0$, it follows from Hölder inequality that

$$
\sup_{k \geq 1} \mathbb{E}^{Q_k} \left[ \sup_{|t-s| \leq \delta, \ 0 \leq s,t \leq T} |\theta_k(t) - \theta_k(s)|^2 \right] \leq \sup_{k \geq 1} \mathbb{E}^{Q_k} \left[ \sup_{|t-s| \leq \delta, \ 0 \leq s,t \leq T} |\theta_k'(u)|^2 \right] \leq \frac{\alpha_{N} \delta}{\lambda_1 \wedge \lambda_2}. \quad (A.5)
$$

For any $\epsilon > 0$ and $\delta > 0$, using Chebychev’s inequality, we arrive at

$$
\sup_{k \geq 1} Q_k \circ \theta_k^{-1} \left( \left\{ h \in \mathcal{C}_{m} ; \sup_{|t-s| \leq \delta, \ 0 \leq s,t \leq T} |h(t) - h(s)| > \epsilon \right\} \right) = \sup_{k \geq 1} Q_k \left( \sup_{|t-s| \leq \delta, \ 0 \leq s,t \leq T} |\theta_k(t) - \theta_k(s)| > \epsilon \right)
\leq \frac{1}{\epsilon^2} \sup_{|t-s| \leq \delta, \ 0 \leq s,t \leq T} |\theta_k(t) - \theta_k(s)|^2 \leq \frac{\alpha_{N} \delta}{(\lambda_1 \wedge \lambda_2) \epsilon^2}. \quad (A.6)
$$

This implies that, for any $\epsilon > 0$,

$$
\lim_{\delta \to 0} \sup_{k \geq 1} Q_k \circ \theta_k^{-1} \left( \left\{ h \in \mathcal{C}_{m} ; \sup_{|t-s| \leq \delta, \ 0 \leq s,t \leq T} |h(t) - h(s)| > \epsilon \right\} \right) = 0. \quad (A.7)
$$

Using similar estimates as (A.5), we obtain that, for all $s \in [0,T]$,

$$
\mathbb{E}^{Q_k} \left[ \sup_{t \in [0,T]} |\theta_k(t)|^2 \right] \leq 2 \mathbb{E}^{Q_k} \left[ |\theta_k(s)|^2 \right] + 2T \mathbb{E}^{Q_k} \left[ \int_{0}^{T} |\theta_k'(u)|^2 du \right], \quad \forall \ k \geq 1. \quad (A.8)
$$

Integrating both sides of the above equation w.r.t. $s$, it follows from Fubini’s theorem that

$$
\mathbb{E}^{Q_k} \left[ \sup_{t \in [0,T]} |\theta_k(t)|^2 \right] \leq \frac{2}{T} \mathbb{E}^{Q_k} \left[ \int_{0}^{T} |\theta_k(s)|^2 ds \right] + 2T \mathbb{E}^{Q_k} \left[ \int_{0}^{T} |\theta_k'(u)|^2 du \right], \quad \forall \ k \geq 1. \quad (A.9)
$$

Then, for any $M > 0$, we obtain from (A.4) that

$$
\sup_{k \geq 1} Q_k \circ \theta_k^{-1} \left( \left\{ h \in \mathcal{C}_{m} ; \sup_{t \in [0,T]} |h(t)| > M \right\} \right) = \sup_{k \geq 1} Q_k \left( \sup_{t \in [0,T]} |\theta_k(t)| > M \right)
\leq \frac{1}{M^2} \sup_{k \geq 1} \mathbb{E}^{Q_k} \left[ \sup_{t \in [0,T]} |\theta_k(t)|^2 \right] \leq \frac{1}{M^2} \left( 2T + \frac{2}{T} \right) \frac{\alpha_{N}}{\lambda_1 \wedge \lambda_2}. \quad (A.10)
$$

By virtue of Arzelà-Ascoli theorem (see, e.g. Simon (1987)) together with (A.7) and (A.10), we obtain that $(Q_k \circ \theta_k^{-1})_{k=1}^{\infty}$, viewed as a sequence of probability measures in $\mathcal{P}(\mathcal{C}_{m})$, is tight.

Note that $\Omega_{\infty} \subset \tilde{\Omega}_{\infty}$ where $\tilde{\Omega}_{\infty} := \Xi_{K_{p}}^{N} \times \mathcal{C}_{p}^{N} \times \mathcal{C}_{m}$. We next claim that the sequence of probability measures $(Q_k \circ \chi_{k}^{-1})_{k=1}^{\infty} \subset \mathcal{P}(\tilde{\Omega}_{\infty})$ is tight. Since we have proven above that $(Q_k \circ \theta_k^{-1})_{k=1}^{\infty} \subset \mathcal{P}(\mathcal{C}_{m})$, it suffices to show that $(Q_k \circ \zeta_{k}^{-1})_{k=1}^{\infty} \subset \mathcal{P}(\Xi_{K_{p}}^{N})$ and $(Q_k \circ W_{k}^{-1})_{k=1}^{\infty} \subset \mathcal{P}(\mathcal{C}_{p}^{N})$ are respectively tight. Note that $(Q_k)_{k=1}^{\infty} \subset \mathcal{Q}(\nu)$, then $Q_k \circ \zeta_k^{-1} = \nu$ for $k \geq 1$. Hence, the tightness of $(Q_k \circ \zeta_k^{-1})_{k=1}^{\infty}$ follows from the fact that $(\mathcal{C}_{p}^{N}, d_{1})$ is Polish. Similarly, it follows from (ii) of Definition 3.1 that the r.v.s $W_k$, for $k \geq 1$, have the same distribution. Thus, $Q_1 \circ W_1^{-1} = Q_k \circ W_k^{-1}$ for all $k \geq 1$, and hence $(Q_k \circ W_k^{-1})_{k=1}^{\infty} \subset \mathcal{P}(\mathcal{C}_{p}^{N})$ is tight, because $(\mathcal{C}_{p}^{N}, d_{2})$ is Polish. By Prokhorov’s theorem, there exists a $Q^* \in \mathcal{Q}(\nu)$ such that $Q_k \circ \chi_k^{-1}$ converges to (up to a subsequence) $Q^*$ in the weak topology of probability measures. Using the Skorokhod’s representation theorem, there exists a probability space $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$, $\Omega_{\infty}$-valued r.v.s $X_k^* = (\zeta_k^*, W_k^*, \theta_k^*)$ with $X_k^* \overset{d}{=} X_k$, and $\mathcal{X} = (\zeta^*, W^*, \theta^*)$ with $\mathbb{P}^* \circ (\mathcal{X})^{-1} = Q^*$ such that, $\mathbb{P}^*$-a.s., as $k \to \infty$,

$$
\zeta_k^* \to \zeta^* \text{ in } \Xi_{K_{p}}^{N}; \quad W_k^* \to W^* \text{ in } \mathcal{C}_{p}^{N}; \quad \theta_k^* \to \theta^* \text{ in } \mathcal{C}_{m}. \quad (A.11)
$$
Let $D := C_0^\infty(\mathbb{R}^m)$ be the space of test functions with its dual space given by $D'$. Denote by $(\cdot, \cdot)$ the dual pair between $D$ and $D'$. We next prove that $\theta_k^\ast$ is $H^1_m$-valued. For any $\phi \in D$, we define the linear functional $T_k(\phi) := (\theta_k^\ast, \phi')$ on $D$, for $k \geq 1$. Let $(\phi)_{i=1}^\infty \subset D$ be dense in $L^2_m$ and consider

$$
\|T_k\| := \sup_{i \geq 1} \frac{|T_k(\phi_i)|}{\|\phi_i\|_{L^2_m}} = \sup_{i \geq 1} \frac{|(\theta_k^\ast, \phi_i')|}{\|\phi_i\|_{L^2_m}}.
$$

Let $E^*$ be the expectation operator under $P^*$. Then, for all $k \geq 1$, it holds that

$$
E^* \left[\sup_{i \geq 1} \frac{|(\theta_k^\ast, \phi_i')|}{\|\phi_i\|_{L^2_m}}\right] = E \left[\sup_{i \geq 1} \frac{|(\theta_k^\ast, \phi_i')|}{\|\phi_i\|_{L^2_m}}\right] = E \left[\|\theta_k^\ast\|_{L^2_m}\right] < +\infty,
$$

and hence $\|T_k\| < +\infty$, $P^*$-a.s. Then, by Hahn-Banach theorem, it holds $P^*$-a.s. that $T_k$ can be extended to be a bounded linear functional on $L^2_m$. Thus, Riesz representation theorem yields the existence of an $L^2_m$-valued random variable $\theta_k^*$ such that $T_k(\phi) = (\theta_k^*, \phi)$ for all $\phi \in L^2_m$, $P^*$-a.s. In particular, $T_k(\phi) = (\theta_k^*, \phi) = (\theta_k^*, \phi)$ for any $\phi \in D$. Hence, $\theta_k^* \in H^1_m$, $P^*$-a.s. It follows that $P^* \circ (X_k^\ast)^{-1} \in P(\Omega_\infty)$. Moreover, we obtain from (A.4) that

$$
\sup_{k \geq 1} E^* \left[\|\theta_k^\ast\|_{L^2_m}\right] = \sup_{k \geq 1} E \left[\sup_{i \geq 1} \frac{|(\theta_k^\ast, \phi_i')|}{\|\phi_i\|_{L^2_m}}\right] = \sup_{k \geq 1} E \left[\|\theta_k^\ast\|_{L^2_m}\right] < \infty.
$$

This implies that $(\theta_k^\ast)_{k=1}^\infty$ is bounded in $L^2((0, T) \times \Omega^*; dt \otimes dP^*)$, and hence $\theta_k^\ast$ (up to a subsequence) converges weakly to some element $\theta^\ast \in L^2((0, T) \times \Omega^*; dt \otimes dP^*)$ as $k \to \infty$. Let $\phi \in D$. By (A.11), we have that, for all $H \in L^\infty(\Omega^*; P^*)$,

$$
E^* [(\theta^\ast, \phi')H] = \lim_{k \to \infty} E^* [(\theta_k^\ast, \phi')H] = -\lim_{k \to \infty} E^* [(\theta_k^\ast, \phi)H] = -E^*[(\theta^\ast, \phi)H].
$$

Then $(\theta^\ast, \phi') = -\langle \theta^\ast, \phi \rangle$, $P^*$-a.s. Therefore, using the separability of $D$, it holds $P^*$-a.s. that $(\theta^\ast, \phi') = -\langle \theta^\ast, \phi \rangle$ for all $\phi \in D$. This gives that $\theta^\ast = \theta^\ast$, $P^*$-a.s. Thus, $P^* \circ (X^\ast)^{-1} \in P(\Omega_\infty)$.

For $k \geq 1$ and $i \in \mathbb{N}$, let $X_k^{*, i}$ be the strong solution of SDE (5) driven by $(\zeta_k^i, W_k^\ast, \theta_k^*)$. In other words, under $(\Omega^*, \mathcal{F}^*, P^*)$, $(X_k^{*, i}(0), Y_k^{*, i}(0)) = \zeta_k^{*, i}$, and for $t \in (0, T]$,

$$
dX_k^{*, i}(t) = f \left(t, \theta_k^*(t), X_k^{*, i}(t), \frac{1}{N} \sum_{j=1}^N \rho(X_k^{*, j}(t))\right) dt + \varepsilon^i dW_k^{*, i}(t).
$$

Moreover, for $i \in \mathbb{N}$, let $X^{*, i}$ be the strong solution of SDE (5) driven by $(\zeta^*, W^*, \theta^*)$. Under $(\Omega^*, \mathcal{F}^*, P^*)$, $(X^{*, i}(0), Y^{*, i}(0)) = \zeta^{*, i}$, and for $t \in (0, T]$,

$$
dX^{*, i}(t) = f \left(t, \theta^*(t), X^{*, i}(t), \frac{1}{N} \sum_{j=1}^N \rho(X^{*, j}(t))\right) dt + \varepsilon^i dW^{*, i}(t).
$$

By the assumption (A., f, p), we obtain that, for all $t \in [0, T]$, $P^*$-a.s.

$$
\sup_{s \in [0, t]} \left|X_k^{*, i}(s) - X^{*, i}(s)\right|^2 \leq C_T \left\{ \left|X_k^{*, i}(0) - X^{*, i}(0)\right|^2 + \int_0^t \left|\theta_k^*(s) - \theta^*(s)\right|^2 ds + \int_0^t \left|X_k^{*, i}(s) - X^{*, i}(s)\right|^2 ds \right. \\
\left. + \int_0^t \left|\frac{1}{N} \sum_{j=1}^N \left(\rho(X_k^{*, j}(s)) - \rho(X^{*, j}(s))\right)\right|^2 ds + |\varepsilon^i| \sup_{s \in [0, t]} \left|W_k^{*, i}(s) - W^{*, i}(s)\right|^2 \right\}.
$$

Using the Lipschitz property of $\rho : \mathbb{R}^d \to \mathbb{R}$, we obtain from Jensen’s inequality that

$$
\left|\frac{1}{N} \sum_{j=1}^N \left(\rho(X_k^{*, j}(t)) - \rho(X^{*, j}(t))\right)\right|^2 \leq \|\rho\|^2_{\text{Lip}} \frac{1}{N} \sum_{j=1}^N \left|X_k^{*, j}(t) - X^{*, j}(t)\right|^2.
$$
Define $|\tilde{x}|_N^2 := \frac{1}{N} \sum_{i=1}^N |x|^2$ for $\tilde{x} = (x^1, \ldots, x^N) \in (\mathbb{R}^d)^N$. It holds that, $\mathbb{P}^*$-a.s.,

$$
\sup_{s \in [0, t]} \left| \tilde{X}_k^*(s) - \tilde{X}^*(s) \right|_N^2 \leq e^{C_r N} \left\{ \left| \tilde{X}_k^*(0) - \tilde{X}^*(0) \right|_N^2 + \int_0^t |\theta_k^*(s) - \theta^*(s)|^2 \, ds + \sup_{s \in [0, t]} |\tilde{W}_k^*(s) - \tilde{W}^*(s)|^2 \right\}.
$$

Therefore, using the convergence results from (A.11), we conclude that, as $k \to \infty$,

$$
\sup_{s \in [0, t]} \left| \tilde{X}_k^*(s) - \tilde{X}^*(s) \right|_N^2 \to 0, \quad \mathbb{P}^*$-a.s. \quad (A.20)
$$

From (A.11) and (A.20), it follows that, $\mathbb{P}^*$-a.s., as $k \to \infty$,

$$
\alpha \left| \tilde{X}_k^*(T) - \tilde{Y}_k^*(0) \right|_N^2 + \beta \int_0^T \left| \tilde{X}_k^*(t) - \tilde{Y}_k^*(0) \right|_N^2 \, dt + \lambda_1 \int_0^T |\theta_k^*(t)|^2 \, dt \\
\quad \to \alpha \left| \tilde{X}^*(T) - \tilde{Y}^*(0) \right|_N^2 + \beta \int_0^T \left| \tilde{X}^*(t) - \tilde{Y}^*(0) \right|_N^2 \, dt + \lambda_1 \int_0^T |\theta^*(t)|^2 \, dt. \quad (A.21)
$$

It follows from properties of convex functionals and weak convergence (see, e.g. Theorem 1.4 in De Figueiredo (1991)) that

$$
\mathbb{E}^* \left[ \|\theta^*\|_{L^2}^2 \right] \leq \liminf_{k \to \infty} \mathbb{E}^* \left[ \|\theta_k^*\|_{L^2}^2 \right]. \quad (A.22)
$$

Recall the sampled objective functional given in (14), and the square form of the loss function and regularizer given in (11). Note that $Q_k = \mathbb{P}^* \circ (\mathcal{X}_k^*)^{-1}$, and $\mathbb{P}^* \circ (\mathcal{X}_k^*)^{-1} = Q^*$. Then, it follows from (A.22) that $Q^* \in \mathcal{Q}(\nu)$. Moreover, by Fatou’s lemma, it follows that

$$
J_N(Q^*) = J_N(\mathbb{P}^* \circ (\mathcal{X}_k^*)^{-1}) \leq \liminf_{k \to \infty} J_N(\mathbb{P}^* \circ (\mathcal{X}_k^*)^{-1}) = \liminf_{k \to \infty} J_N(Q_k). \quad (A.23)
$$

We then deduce that $J_N(Q^*) \leq \alpha_N$ by using (A.1). Recall that $Q^* \in \mathcal{Q}(\nu)$ and hence $\alpha_N \leq J_N(Q^*)$. Therefore $J_N(Q^*) = \alpha_N$, i.e., $Q^* \in \mathcal{Q}(\nu)$ is the optimal relaxed solution of (15). This ends the proof. \qed

To prove Lemma 4.2 and Lemma 4.3, we need the following technical lemma. Its proof is omitted because it is standard, and based on an application of the Itô’s formula and use of Doob’s maximal inequality (the proof can be provided upon request).

**Lemma A.1.** Let $\tilde{X}_N(t) = (X^1_N(t), \ldots, X^N_N(t))$ for $t \in [0, T]$ with $X^i_N(t)$ satisfying SDE (19). Let $(A_{\varepsilon, f, \rho})$ and $(A_{d})$ hold. Then, for all $p \geq 1$, there exists a constant $C_p$ which is independent of $N$ such that

$$
\mathbb{E}^{Q_N} \left[ \sup_{t \in [0, T]} \left| \tilde{X}_N(t) \right|_N^{2p} \right] \leq C_p, \quad \forall \ N \geq 1. \quad (A.24)
$$

**Proof of Lemma 4.2.** It follows from (21) and Chebyshev’s inequality that

$$
\sup_{N \geq 1} \mathbb{E}^{Q_N} \left[ \left\{ \varepsilon \in \mathbb{E}_N \right\} \sup_{t \in [0, T]} \left| e^{2 \varepsilon \theta(t, \varepsilon)} - M \right| \right] \leq \frac{1}{M} \sup_{N \geq 1} \mathbb{E}^{Q_N} \left[ \sup_{t \in [0, T]} \left| e^{2 \varepsilon \theta(t, \varepsilon)} - M \right| \right] \leq \frac{1}{M} \sup_{N \geq 1} \mathbb{E}^{Q_N} \left[ \sup_{t \in [0, T]} \left| e^{2 \varepsilon \theta(t, \varepsilon)} - M \right| \right] \leq C_p \left\{ \sup_{N \geq 1} \mathbb{E}^{Q_N} \left[ \left| \xi_N^i \right|^{2p} \right] + \sup_{N \geq 1} \mathbb{E}^{Q_N} \left[ \sup_{t \in [0, T]} \left| \tilde{X}_N(t) \right|_N^{2p} \right] \right\}. \quad (A.25)
$$

By the assumption $(A_{\varepsilon, f, \rho})$-(i) and noting that $\xi^i \in \Xi_K$ for all $i \geq 1$, the limit (26) follows from Lemma A.1. Using the representation of the empirical measure given by (20), it holds that, for $s, t \in [0, T],$

$$
W^2_{E, 2} (\mu^N(t), \mu^N(s)) \leq \frac{1}{N} \sum_{i=1}^N \left| X^i_N(t) - X^i_N(s) \right|^2. \quad (A.26)
$$
Then, it follows from Chebyshev’s inequality that
\[
\sup_{N \geq 1} Q^N \left( \left\{ \theta \in \hat{S}: \sup_{|t-s| \leq \delta} W_{E,2}(\theta(t), \theta(s)) > \varepsilon \right\} \right) \leq \frac{1}{\varepsilon^2} \sup_{N \geq 1} \mathbb{E}^{Q_N} \left[ \frac{1}{N} \sum_{i=1}^{N} \sup_{|t-s| \leq \delta} |X^i_N(t) - X^i_N(s)|^2 \right].
\]
Using the assumption (A.27) and Lemma A.1 with the assumption (A_\theta), we deduce the existence of a positive constant C which is independent of N such that
\[
\sup_{N \geq 1} \mathbb{E}^{Q_N} \left[ \frac{1}{N} \sum_{i=1}^{N} \sup_{|t-s| \leq \delta} |X^i_N(t) - X^i_N(s)|^2 \right] \leq \delta^2 \{1 + \Upsilon(\delta)\}, \quad (A.27)
\]
where, for \(\delta > 0\), we have defined
\[
\Upsilon(\delta) := \sup_{N \geq 1} \mathbb{E}^{Q_N} \left[ \frac{1}{N} \sum_{i=1}^{N} \sup_{|t-s| \leq \delta} |W^i_N(t) - W^i_N(s)|^2 \right]. \quad (A.28)
\]
Note that \((W^1_N, \ldots, W^N_N)\) are independent Wiener processes under \(Q_N\). Hence, Doob’s maximal inequality implies
\[
\Upsilon(\delta) = \sup_{N \geq 1} \mathbb{E}^{Q_N} \left[ \frac{1}{N} \sum_{i=1}^{N} \sup_{|t-s| \leq \delta} |W^i_N(t) - W^i_N(s)|^2 \right] = \sup_{N \geq 1} \mathbb{E}^{Q_N} \left[ \sup_{|t-s| \leq \delta} |W^i_N(t) - W^i_N(s)|^2 \right] \leq 4\delta. \quad (A.29)
\]
This yields \(\Upsilon(\delta) \to 0\) as \(\delta \to 0\). Hence, the limit (27) follows from (A.27).

**Proof of Lemma 4.8.** Define the mappings \(\hat{I}_N, \hat{I}_* : \Xi^N_R \to \Omega^0_{\infty} \times \mathcal{P}_2(E)\) as follows: for any \(\hat{\zeta} \in \Xi^N_R, \)
\[
\hat{I}_N(\hat{\zeta}) := (\hat{\zeta}, I_N(\hat{\zeta})), \quad \hat{I}_*(\hat{\zeta}) := (\hat{\zeta}, I_*(\hat{\zeta})). \quad (A.30)
\]
It follows from (58) in (A.4) that \(\nu \circ \hat{I}_N^{-1} \nu \circ \hat{I}_N^{-1} \text{ as } N \to \infty\). Observe that \(Q_N \circ \hat{I}_N(\zeta_N)^{-1} = \{Q_N \circ \zeta_N^{-1}\} \circ \hat{I}_N^{-1} = \nu \circ \hat{I}_N^{-1}, \text{ and } Q \circ \hat{I}_*(\zeta)^{-1} = \nu \circ \hat{I}_*^{-1}. \) Then
\[
Q_N \circ \hat{I}_N(\zeta_N)^{-1} \Rightarrow Q \circ \hat{I}_*(\zeta)^{-1} \text{ as } N \to \infty. \quad (A.31)
\]
Using the inequality \(\mathcal{W}_{\Omega,\infty,2}(Q_N \circ (\zeta_N, \theta_N)^{-1}, Q \circ (\zeta, \theta)^{-1}) \leq \mathcal{W}_{\Omega,\infty,2}(Q_N, Q)\) and the assumption that \(\lim_{N \to \infty} \mathcal{W}_{\Omega,\infty,2}(Q_N, Q) = 0\), we arrive at
\[
\lim_{N \to \infty} \mathcal{W}_{\Omega,\infty,2}(Q_N \circ (\zeta_N, \theta_N)^{-1}, Q \circ (\zeta, \theta)^{-1}) = 0. \quad (A.32)
\]
Combining (A.31) and (A.32), we obtain that \((Q_N \circ (\zeta_N, I_N(\zeta_N), \theta_N)^{-1})_{N=1}^\infty\) is tight.

We next prove that any convergent subsequence of \((Q_N \circ (\zeta_N, I_N(\zeta_N), \theta_N)^{-1})_{N=1}^\infty\) has the same weak limit. To start with, let \((N^i_{k})_{k=1}^\infty, i = 1, 2,\) be two subsequences of \(N\) such that \(Q_{N^i_k} \circ (\zeta_{N^i_k}, I_{N^i_k}(\zeta_{N^i_k}), \theta_{N^i_k})^{-1} \Rightarrow \mathbb{P} \circ (\zeta_i, J_i, \theta_i)^{-1} \text{ as } k \to \infty.\) Here, \((\zeta_i, J_i, \theta_i)\) is an \(\Omega^0_{\infty} \times \mathcal{P}_2(E) \times \mathcal{H}^1_{\infty}\)-valued random variable defined on some probability space \((\Omega^i, \mathcal{F}^i, \mathbb{P}^i)\). By (A.31), we have that \(\mathbb{P}^i \circ (\zeta_i, J_i)^{-1} = \nu \circ \hat{I}_i^{-1} \text{ for } i = 1, 2.\) It then follows from (A.30) that, for \(i = 1, 2,\)
\[
\mathbb{P}^i \left( \{\omega \in \Omega^i: J_i = I_*(\zeta_i(\omega))\} \right) = \nu \left( \{\hat{\zeta} \in \Xi^N_R: \hat{I}_*(\hat{\zeta}) = (x_1, x_2) = I_*(x_1)\} \right) = 1. \quad (A.33)
\]
By applying the Gluing lemma (see Lemma 7.6 in Villani (2003)), there exists a coupling \((J_1^*, J_2^*, \zeta^*, \theta^*)\) under some probability space \((\Omega^*, \mathcal{F}^*, \mathbb{P}^*)\) such that \((\zeta^*, J_1^*, \theta^*) = (\zeta_i, J_i, \theta_i)\) in law for \(i = 1, 2.\) It then follows from (A.33) that \(J_1^* = J_2^* = I_*(\zeta^*), \mathbb{P}^*\)-a.s. This yields that \(\mathbb{P}^1 \circ (\zeta_1, J_1, \theta_1)^{-1} = \mathbb{P}^2 \circ (\zeta_2, J_2, \theta_2)^{-1}, \) and hence every convergent subsequence of \((Q_N \circ (\zeta_N, I_N(\zeta_N), \theta_N)^{-1})_{N=1}^\infty\) admits the same weak limit. Moreover, for \(Q \in \mathcal{Q}(\nu),\) the assumption (A_\nu) together with Definition 3.1-(i) yields
\[
Q \left( \{\omega \in \Omega^\infty: \lim_{N \to \infty} W_{E,2}(I_N(\zeta_N(\omega)), I_*(\zeta_N(\omega))) = 0 \} \right) = \nu \left( \{\hat{\zeta} \in \Xi^N_R: \lim_{N \to \infty} W_{E,2}(I_N(\hat{\zeta}), I_*(\hat{\zeta})) = 0 \} \right) = 1,
\]
and hence \((\zeta_N, I_*(\zeta_N), \theta) \to (\zeta, I_*(\zeta), \theta), \) Q-a.s. This concludes the proof that \(Q \circ (\zeta_N, I_N(\zeta_N), \theta_N)^{-1} \Rightarrow Q \circ (\zeta, I_*(\zeta), \theta)^{-1}\) as \(N \to \infty.\) Moreover, define a specific sequence \((\tilde{Q}_N)_{N=1}^\infty \subset \mathcal{Q}(\nu)\) as \(\tilde{Q}_{2l-1} := Q_{2l-1} \text{ and } \tilde{Q}_{2l} = Q \text{ for all } l \in \mathbb{N}.\) We then have that \(\tilde{Q}_N \circ (\zeta_N, I_N(\zeta_N), \theta_N)^{-1} \Rightarrow Q \circ (\zeta, I_*(\zeta), \theta)^{-1}.\) This proves the weak convergence result in the lemma.

\[\square\]
Proof of Lemma 4.10. The uniqueness of a solution to the FPK equation (63) in the trajectory sense follows from Proposition 4.4. We next show the existence. For given $(I_*, \theta) \in \mathcal{P}_2(E) \times \mathcal{C}_m$, consider the weak solution of the following parameterized SDE defined on a filtered probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}} = (\hat{\mathcal{F}}_t)_{t \in [0,T]}, \hat{\mathbb{P}})$ which supports a $p$-dimensional Brownian motion $\hat{W} = (\hat{W}(t))_{t \in [0,T]}$, and r.v.s $(Y(0), X(0)) \in \hat{\mathcal{F}}_0$:

$$dX^\varepsilon,Y,\theta(t) = f(t, \theta(t), X^\varepsilon,Y,\theta(t)), d\hat{W}(t),$$

(A.34)

and $(\varepsilon, Y(0), X(0))$ admits the law $I_*$. It can then be verified that $\mu(t) := \hat{\mathbb{P}} \circ (\varepsilon, Y(0)), X^\varepsilon,Y,\theta(t))^{-1}$, for $t \in [0,T]$, satisfies Eq. (63).

Moreover, by Lemma 4.8, $Q_N \circ (\mu^N_0, \theta_N)^{-1} \Rightarrow Q \circ (I_*, \theta)^{-1}$ as $N \to \infty$. It follows from Theorem 4.1 that $Q_N := Q_N \circ (\mu^N_0, \theta_N, \mu^N) \to \hat{\mathbb{P}} \circ (\mu_0, \theta, \hat{\mu})$ in $\mathcal{P}_2(\mathcal{P}_2(E) \times \mathcal{C}_m \times S)$ for some probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$, where $\hat{\mathbb{P}}$-a.s., $\hat{\mu}$ is the unique solution of FPK equation (24) with initial condition $\hat{\mu}(0) = \mu_0$. Note that $(\hat{\mu}_0, \theta)$ has the law given by $Q \circ (I_*, \theta)^{-1}$. Then, from Gluing lemma, there exists a coupling $(\hat{\mu}_0, \theta, \hat{\mu}^1, \hat{\mu}^2)$ on some probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$ such that $\hat{\mathbb{P}} \circ (\hat{\mu}_0, \theta, \hat{\mu})^{-1} = \hat{\mathbb{P}} \circ (\hat{\mu}_0, \theta, \hat{\mu}^1)^{-1}$ and $\hat{\mathbb{P}} \circ (\hat{\mu}_0, \theta, \hat{\mu}^2)^{-1} = Q \circ (I_*, \theta, \mu_*)^{-1}$. Recall here that, $Q$-a.s., $\mu_*$ solves FPK equation (24) with initial condition $\mu_*(0) = I_*$. Using a similar proof to that of Lemma 4.5, it follows that $(\hat{\mu}_0, \theta, \hat{\mu})$ and $(I_*, \theta, \mu_*)$ are identical in law. Taking $Q_N \Rightarrow Q$ for all $N \geq 1$, we have that $\hat{Q}^* = Q \circ (I_*, \theta, \mu_*)^{-1}$.

References

Ahmed, N.U., and C.D. Charalambous (2013): Stochastic minimum principle for partially observed systems subject to continuous and jump diffusion processes and driven by relaxed controls. SIAM J. Contr. Optim. 51, 3235-3257.

Aliprantis, C.D., and K. Border (2006): Infinite Dimensional Analysis: A Hitchhiker’s Guide. Springer-Verlag, New York.

Annunziato, M., and A. Borzì (2010): Optimal control of probability density functions of stochastic processes. Math. Modelling Anal. 15, 393-407.

Ba, J., and B. Frey (2013): Adaptive dropout for training deep neural networks. In Advances in Neural Information Processing Systems 26, 1-9.

Bahalli, S., B. Djehiche, and B. Mezerdi (2007): The relaxed stochastic maximum principle in singular optimal control of diffusions. SIAM J. Contr. Optim. 46, 427-444.

Carmona, R., F. Delarue, and D. Laccker (2016): Mean field games with common noise. Ann. Probab. 44, 3740-3803.

Chen, R., Y. Rubanova, J. Bettencourt, and D. Duvenaud (2018): Neural ordinary differential equations. In Adv. in Neural Infor. Processing Syst., 6571-6583.

Chizat, L. and F. Bach (2018): On the global convergence of gradient descent for over-parameterized models using optimal transport. Advances in Neural Information Processing Systems (NeurIPS), 3040-3050.

Cuchiero, C., M. Larsson, and J. Teichmann (2020): Deep neural networks, generic universal interpolation, and controlled ODEs. SIAM J. Math. Data Sci. 2, 901-919.

Dal Maso, G. (1993): An Introduction to Γ-Convergence. Springer-Verlag, New York.

De Figueiredo, D.G. (1991): Lectures on the Ekeland variational principle with applications and detours. Acta. Appl. Math. 24, 195-196.

Dieng, A.B., R. Ranganath, J. Altosaar, and D.M. Blei (2018): Noising: Unbiased Regularization for Recurrent Neural Networks. Proceeding of the 35th International Conference on Machine Learning 80, 1252-1261.

Du, S.S., J. Lee, H. Li, L. Wang, and X. Zhai (2019): Gradient descent finds global minima of deep neural networks. Proceeding of the 35th International Conference on Machine Learning, 3003-3048.
E, Weinan (2017): A proposal on machine learning via dynamical systems. *Commun. Math. Stats.* 5, 1-11.

E, Weinan, J.Q. Han, and Q.X. Li (2018): A mean-field optimal control formulation of deep learning. *Res. Math. Sci.* 6:10, 1-41.

El Karoui, N., D.H. Nguyen, and M. Jeanblanc-Picquè (1987): Compactification methods in the control of degenerate diffusions: existence of an optimal control. *Stochastics* 20, 169-219.

Evans, L.C. (2010): *Partial Differential Equations, 2nd Ed.* AMS, Providence.

Jabir, J.F., D. Siska, and L. Szpruch (2019): Mean-field neural ODEs via relaxed optimal control. Preprint, available at https://arxiv.org/abs/1912.05475

Hasan, M., and A.K. Roy-Chowdhury (2015): A continuous learning framework for activity recognition using deep hybrid feature models. *IEEE Trans. Multimedia* 17, 1909-1922.

Haussmann, U.G., and J.P. Lepeltier (1990): On the existence of optimal controls. *SIAM J. Contr. Optim.* 28, 851-902.

Haykin, S. (2009): *Neural Network and Learning Machines, 3rd Ed.* Pearson Education Inc, New York.

He, K., S. Ren, and J. Sun (2016): Deep residual learning for image recognition. In *Proceedings of IEEE Confer. Comput. Vis. Pattern Recogn.* 770-778.

Hintermüller, M., D. Marahrens, P.A. Markowich, and C. Sparber (2013): Optimal bilinear control of Gross-Pitaevskii equations. *SIAM J. Contr. Optim.* 51, 2509-2543.

Ioffe, S., and C. Szegedy (2015): Batch normalization: accelerating deep network training by reducing internal covariate shift. *ICML'15 Proceedings of the 32nd International Conference on Machine Learning* 37, 448-456.

Kingma, D.P., T. Salimans, and M. Welling (2015): Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems* 28, 2575-2583.

Lacker, D. (2015): Mean field games via controlled martingale problems: Existence of Markovian equilibria. *Stoch. Process. Appl.* 125, 2856-2894.

Lacker, D. (2016): A general characterization of the mean field limit for stochastic differential games. *Probab. Theor. Relat. Fields* 165, 581-648.

Lu, Y., A. Zhong, Q. Li, and -B. Dong (2018): Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *International Conference on Machine Learning*, 5181-5190.

Mallat, S. (2016): Understanding deep convolutional neural networks. *Philos Trans. A. Math. Phys. Eng. Sci.* 374, 20150203.

Manita, O.A., M.S. Romanov, and S.V. Shaposhnikov (2015): On uniqueness of solutions to nonlinear Fokker-Planck-Kolmogorov equations. *Nonlinear Anal.* 128, 199-226.

Mei, S., A. Montanari, and P. Nguyen (2018): A mean field view of the landscape of two-layer neural networks. *PNAS* 115, E7665-E7671.

Rotskoff, G.M., and E. Vanden-Eijnden (2019): Trainability and accuracy of neural networks: An interacting particle system approach. Preprint, available at https://arxiv.org/abs/1805.00915v3

Simon, J. (1987): Compact sets in the space $L^p(0,T;B)$. *Ann. Mat. Pura Appl.* 146, 65-96.

Sirignano, J. and K. Spiliopoulos (2021): Mean field analysis of neural networks: A law of large numbers. *SIAM J. Appl. Math.* 80, 725-752.

Sirignano, J., and K. Spiliopoulos (2020): Mean field analysis of deep neural networks. *Math. Opers. Res.* Forthcoming.
Srivastava, N., G.E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014): Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929-1958.

Thorpe, M., and Y. van Gennip (2020): Deep limit of residual neural networks. Preprint, available at [https://arxiv.org/abs/1810.11741](https://arxiv.org/abs/1810.11741)

Villani, C. (2003): *Topics in Optimal Transportation.* Graduate Studies in Mathematics, Volume 58, AMS.

Villani, C. (2009): *Optimal Transport: Old and New Part.* Springer-Verlag, New York.

Wan, L., M. Zeiler, S. Zhang, Y. LeCun, and R. Fergus (2013): Regularization of neural networks using dropconnect. *In Intern. Conf. Machine Learn.* 1, 1-12.