A BAYESIAN VARIATIONAL FRAMEWORK FOR STOCHASTIC OPTIMIZATION

PHILIPPE CASGRAIN

Department of Statistical Sciences, University of Toronto

Abstract. This work proposes a theoretical framework for stochastic optimization algorithms, based on a continuous-time Bayesian variational model. Using techniques from stochastic control with asymmetric information, the solution to this variational problem is shown to be equivalent to a system of Forward Backward Differential Equations (FBSDEs). Using an analytical approximation to the solution of these FBSDEs, we recover a variety of existing adaptive stochastic gradient descent methods. This framework establishes a direct connection between stochastic optimization algorithms and a secondary Bayesian inference problem on gradients, where the prior and assumed observation dynamics determine the resulting algorithm.

1. Introduction

Stochastic optimization algorithms are a widely used tools which are crucial to solving to optimization problems arising in machine learning. The initial motivation for algorithms arises from the fact that as the scale and dimension of an optimization problem becomes large, computing the gradients of the target loss function becomes increasingly difficult from a computational standoint. This causes deterministic gradient-based optimization algorithms to perform poorly, due to the increased computational load of repeatedly computing gradients. Stochastic optimization algortihms fill this gap by replacing exact gradients of the target loss with a computationally cheap gradient estimator, trading off noise in gradient estimates for an computational efficiency at each step.

Let us consider the problem of minimizing a generic risk function $f : \mathbb{R}^d \to \mathbb{R}$, taking the form

$$f(x) = \frac{1}{|\mathfrak{N}|} \sum_{z \in \mathfrak{N}} \ell(x; z) \, , \tag{1.1}$$

where $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$, and where we define the set $\mathfrak{N} := \{z_i \in \mathcal{Z} \ , \ i = 1, \dots, N\}$ to be a set of training points. In this definition, we interpret $\ell(x; z)$ as the model loss at a training point $z \in \mathfrak{N}$ for the parameters $x \in \mathbb{R}^d$.

Since N and d are typically large, computing gradients of f at an arbitrary point x can be expensive, and can make the use of deterministic gradient-based algorithms extremely inefficient. Let us consider an optimization algorithm $\{x_t\}_{t \in \mathbb{N}}$. Rather than computing $\nabla f(x_t)$ directly at each point of the optimization

E-mail address: p.casgrain@mail.utoronto.ca.

PC would like to thank Daniel Roy and Sebastian Jaimungal for the fruitful discussions and useful feedback on this work.

process, we may instead collect noisy samples of gradients as

$$g_t = \frac{1}{|\mathfrak{N}_t^m|} \sum_{z \in \mathfrak{N}_t^m} \ell(x; z) \ , \tag{1.2}$$

where for each t, $\mathfrak{N}_t^m \subseteq \mathfrak{N}$ is an independent sample of size m from the set of training points. Typically, we assume that $m \ll N$ is chosen small enough so that computing g_t can be done a a significantly lower cost than computing the full gradient, $\nabla f(x_t)$. Using the stream of samples $\{g_t\}$, stochastic optimiztion algorithms construct an estimate $\widehat{\nabla f}(x_t)$ of the quantity $\nabla f(x_t)$ in order to determine the next step x_{t+1} for the optimizer.

There exists a large body of literature stochastic optimization algorithms. Each algorithm introduces its own variation on the gradient estimator $\widehat{\nabla f}(x_t)$ as well as other features which may enhance convergence to an optimum. Among the simplest of these is stochastic gradient descent and its variants [22], which use an estimator based on single gradient samples. Other examples include momentum methods and its variants, such as [16, 19], which can be interpreted as using exponentially weighted moving averages (EWMA) as gradient estimators. Adaptive gradient descent methods such as AdaGrad [9] or Adam [15] use similar moving average estimators, as well as dynamically updated normalization factors. For a survey paper which covers many of the modern stochastic optimization that are used in practice see [23]. Some previous related works, such as [8], [10] and [1] also describe theories to unify stochastic optimization algorithms. This paper takes instead takes a variational approach, as in Wibisono, Wilson & Jordan [28], who provide a variational model for deterministic optimization algorithms.

This paper interprets the problem of stochastic optimization as equivalent to the problem of optimizing a latent function. More specifically, we assume that optimizers cannot directly observe the loss function $f(x)$. Instead, an optimizer's available set of information consists solely of the collection of noisy gradient samples, $\{g_t\}$, that is accumulated over the course of the optimization. By specifying a variational problem over optimizers with this information restriction feature and by placing a prior measure over possible loss functions, we find that optimizers which are solutions to this variational problem naturally generate bayesian gradient estimators $\widehat{\nabla f}(x_t)$ within their dynamics. Moreover, the optimizer's behaviour is entirely determined by only a handful of model parameters, including the assumed model prior and the assumed model for noisy observations. From this framework, we demonstrate that under simple models, we can recover a number of optimization algorithms, demonstrating that *these are in fact variational.*

This work can be considered most similar to [28], where the variational model for deterministic algortihms is extended to the stochastic case. This is achieved by introducing the additional restriction that optimizers may not directly evaluate gradients of the target loss function, but instead only have access to noisy streams of gradients. By making this notion mathematically rigourous via measurability arguments, we can derive optimality equations taking the form of Forward-Backward Stochastic Differential Equations (FBSDEs). As a result, we obtain algorithms which have a built-in online learning property. More specifically, these algorithms use a Bayesian filter on the noisy gradient stream g_t to compute estimates of $\nabla f(x_t)$. Moreover, we find that under various a-priori model assumptions on ∇f and g, we can recover a number of common stochastic optimization algorithms.

The paper is structured as follows. Section 2 presents a Bayesian surrogate model for stochastic optimization. Section 3 introduces and motivaties a stochastic variational problem over optimizers. Section 4 presents necessary and sufficient conditions for an optimizer to form a solution to the variational

priblem. Lastly, Sections 5 and 6 demonstrate how various discrete optimization algorithms are recovered through discretizations of the continuous optimality equations derived in Section 4.

2. A Statistical Model for Stochastic Optimization

Over the course of the section we lay out the variational model for our optimization problem. The ultimate objective will be to obtain an algorithm to minimize a loss function $f(x)$ over the domain $\mathbb{R}^d$, where f is assumed to be a single draw from a random variable which generates differentiable loss functions.

More precisely, let us consider the ambient probability space $(\Omega, \mathbb{P}, \mathfrak{G} = \{\mathcal{G}_t\}_{t\in[0,T]})$, where $\mathcal{G}_t$ is a filtration which we define over the course of this section. Our model assume that loss function we wish to minimize is a random variable $f : \Omega \to C^1(\mathbb{R}^d)$. Each draw from the random variable $f(x)$ is assumed to be a continuosly differentiable function from the domain $\mathbb{R}^d$ into the real numbers. We make the additional technical assumption that $\mathbb{E}\,\|\nabla f(x)\|^2 < \infty$ for all $x \in \mathbb{R}^d$.

We define the optimizer $X = (X_t^\nu)_{t\geq 0}$ as a controlled continuous process satisfying $X_t^\nu \in \mathbb{R}^d$ for all $t \geq 0$, with an initial condition $X_0 \in \mathbb{R}^d$. We make the assumption that the paths of X are continuously differentiable in time. Thus, we assume that X has the dynamics $dX_t^\nu = \nu_t\, dt$ where $\nu_t \in \mathbb{R}^d$ represents the control, and where we use the superscript to express the explicit dependence of X^ν on the control ν. Hence, we may write the optimizer in its integral form as

$$X_t^\nu = X_0^\nu + \int_0^t \nu_u\, du\ , \tag{2.1}$$

demonstrating that the optimizer is entirely characterized by a pair (ν, X_0) consisting of a control process ν and an initial condition X_0^ν. Using an explicit Euler discretization with step size $\epsilon > 0$, the optimizer can be approximately represented through the update rule $X_{t+\epsilon}^\nu \approx X_t^\nu + \epsilon\,\nu_t$. This leads to the interpretation of ν_t as the (infinitesimal) step the algorithm takes at each point t during the optimization process.

In order to capture the essence of the problem that is present in stochastic optimization, we construct our model so that there is restricted access to the gradients of the loss function f. Rather than being able to directly observe gradients of f over the path of X_t^ν, we assume that we only have access to a noisy source of noisy gradients expressed by the càdlàg semi-martingale $g = (g_t)_{t\geq 0}$. As a simple motivating example, we can consider the model $g_t = \nabla f(X_t^\nu) + \xi_t$, where ξ_t is a white noise process. This observation model can be interpreted as consisting of the true value of $\nabla f(X_t^\nu)$ plus an i.i.d. source of noise. This concrete example will be useful to keep in mind to make sense of some of the results which we present over the course of the paper.

To make this information restriction concept mathematically rigorous, we restrict ourselves only to optimizers X^ν which are measurable with respect to the information generated by the noisy gradient process g. In order to so this, we first define the global filtration $\mathcal{G}$, as $\mathcal{G}_t = \sigma\left((g_u)_{u\in[0,t]}, f\right)$, the sigma algebra generated by the paths of g as well as the realizations of the loss surface f. The filtration $\mathcal{G}_t$ is defined to contain the complete set of information generating the optimization problem until time t.

Next, we define the coarser filtration $\mathcal{F}_t = \sigma(g_u)_{u\in[0,t]} \subset \mathcal{G}_t$ generated strictly by the paths of the noisy gradient process. This filtration represents the total set of information available to the optimizer up until time t. This allows us to formally restrict the flow of information to the algorithm by restricting our choice of possible optimizers to those which are $\mathcal{F}_t$-progressively-measurable. More precisely, we say

that the optimizer's control ν is admissible if

$$\nu \in \mathcal{A} = \left\{ \omega = (\omega_t)_{t\geq 0} \,:\, \omega \text{ is } \mathcal{F}\text{-adapted}\,,\ \mathbb{E}\int_0^T \|\omega_t\|^2 + \|\nabla f(X_t^\omega)\|^2\, dt < \infty \right\}\,. \tag{2.2}$$

The set of optimizers generated by $\mathcal{A}$ can be interpreted as the set of $\mathcal{F}_t$-adapted optimizers which have bounded expected travel distance and square-integrable gradients over their path.

3. The Optimizer's Variational Problem

Now that we have defined the set of admissible optimization algortihms, we set out to select those which are optimal in an appropriate sense. We proceed similarly to Wibisono, Wilson and Jordan [28], by proposing an objective functional which measures the performance of the optimizer over a finite time period.

The motivation for the optimizers performance metric comes from a physical interpretation of the optimization process. We can think of our optimization process as a particle travelling through a potential field define by the target loss function f. As the particle travels through the potential field, it may either gain or lose momentum depending on its location and velocity, which will in turn affect the particle's trajectory. Naturally, we may seek to find the path of a particle which reaches the optimum of the loss function while minimizing the total amount of kinetic and potential energy that is spent. We therefore turn to the Lagrangian interpretation of classical mechanics, which provides a framework for obtaining solutions to this problem. Over the remainder of this section, we lay out the Lagrangian formalism for the optimization problem we defined in Section 2.

To define a notion of energy in the optimization process, we provide a measure of distance in the parameter space. We use the *Bregman Divergence* as the measure of distance within our parameter space, which can embed additional information about the geometry of the optimization problem. The Bregman divergence, D_h, is defined as

$$D_h(y,x) = h(y) - h(x) - \langle \nabla h(x), y - x\rangle \tag{3.1}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is a strictly convex function satisfying $h \in C^2$. We assume here that the gradients of h are L-Lipschitz smooth for a fixed constant $L > 0$. The choice of h determines the way we measure distance, and is typically chosen so that it mimics features of the loss function f. In particular, this quantity plays a central role in mirror descent and non-linear sub-gradient algorithms. For more information on this connection and on Bregman Divergence, see [18] and [2].

We define the total energy in our problem as the kinetic energy, accumulated through the movement of the optimizer, and the potential energy generated by the target loss function f. Under the assumption that f almost surely admits a global minimum $x^\star = \arg\min_{x\in\mathbb{R}^d} f(x)$, we may represent the total energy via the Bregman Lagrangian as

$$\mathcal{L}(t, X, \nu) = e^{\gamma_t}\Big(\underbrace{e^{\alpha_t} D_h\left(X + e^{-\alpha_t}\nu, X\right)}_{\text{Kinetic Energy}} - \underbrace{e^{\beta_t}\left(f(X) - f(x^\star)\right)}_{\text{Potential Energy}} \Big)\,, \tag{3.2}$$

for fixed inputs (t, X, ν), and where we assume that $\gamma, \alpha, \beta : \mathbb{R}^+ \to \mathbb{R}$ are deterministic, and satisfy $\gamma, \alpha, \beta \in C^1$. The functions γ, α, β can be interpreted as hyperparameters which tune the energy present

at any state of the optimization process. An important property to note is that the Lagrangian is itself a random variable due to the randomness introduced by the latent loss function f.

The objective is then to find an optimizer within the admissible set $\mathcal{A}$ which can get close to the minimum $x^\star = \min_{x\in\mathbb{R}^d} f(x)$, while simultaneously minimizing the energy cost over a finite time period $[0,T]$. The approach taken in classical mechanics and in [28] fixes the endpoint of the optimizer at $x^\star$. Since we assume that the function f is not directly visible to our optimizer, it is not possible to add a constraint of this type that will hold almost surely. Instead, we introduce a soft constraint which penalizes the algorithm's endpoint in proportion to its distance to the global minimum, $f(X_T) - f(x^\star)$. As such, we define the expected action functional $\mathcal{J} : \mathcal{A} \to \mathbb{R}$ to be

$$\mathcal{J}(\nu) = \mathbb{E}\Big[\underbrace{\int_0^T \mathcal{L}(t, X_t^\nu, \nu_t)\, dt}_{\text{Total Path Energy}} + \underbrace{e^{\delta_T}\Big(f(X_T^\nu) - f(x^\star)\Big)}_{\text{Soft End Point Constraint}} \Big] , \tag{3.3}$$

where $\delta_T \in C^1$ is assumed to be an additional model hyperparameter, which controls the strength of the soft constraint.

With this definition in place, the objective will be to select amongst admissible optimizers for those which minimize the expected action. Hence, we seek optimizers which solve the stochastic variational problem

$$\nu^* = \arg\min_{\nu\in\mathcal{A}} \mathcal{J}(\nu) \, . \tag{3.4}$$

Remark 3.1. Note that the variational problem (3.4) is identical to the one with Lagrangian

$$\tilde{\mathcal{L}}(t, X, \nu) = e^{\gamma_t}(e^{\alpha_t} D_h\left(X + e^{-\alpha_t}\nu, X\right) - e^{\beta_t} f(X)) \tag{3.5}$$

and terminal penalty $e^{\delta_T} f(X_T^\nu)$, since they differ by constants independent of ν. Because of this, the results presented in Section 4 also hold the case where $x^\star$ and $f(x^\star)$ do not exist or are infinite.

The optimization problem (3.4) presents a number of technical difficulties which prevent us from passing it through the standard machinery of Lagrangian and Hamiltonian mechanics or of stochastic control. The main hurdle stems from the measureability restriction of ν to $\mathcal{F}_t$, which introduces an asymmetry between the control's filtration $\mathcal{F}_t$ and the filtration generating dynamics of the optimization problem, $\mathcal{G}_t$.

4. Critical Points of the Expected Action Functional

In order to solve the variational problem (3.4), we make use techniques from the calculus of variations and infinite dimensional convex analysis to provide optimality conditions for minima of the expected action functional. To address issues of information asymmetry, we adapt the stochastic control techniques of Casgrain & Jaimungal [5–7], originally developed in the context of stochastic mean-field games with incomplete information.

The approach we take relies on the fact that a necessary condition for the optimality of a Gâteaux differentiable functional $\mathcal{J}$ is that its Gâteaux derivative vanishes in all appropriate directions. Computing the Gâteaux derivative of $\mathcal{J}$, we find an equivalence between the Gâteaux derivative vanishing and a system of Forward-Backward Stochastic Differential Equations (FBSDEs), yielding a generalization of the Euler-Lagrange equations to the context of our optimization problem. The precise result is stated in Theorem 4.1 below.

Theorem 4.1 (Stochastic Euler-Lagrange Equation)**.** *A control $\nu^* \in \mathcal{A}$ is a critical point of $\mathcal{J}$ if and only if (X^{ν^*}, ν^*) forms a solution to the system of FBSDEs,*

$$\begin{cases} d\left(\frac{\partial \mathcal{L}}{\partial \nu}\right)_t = \mathbb{E}\left[\left(\frac{\partial \mathcal{L}}{\partial X}\right)_t \Big| \mathcal{F}_t\right] dt + d\mathcal{M}_t \ , \quad t < T \\ \left(\frac{\partial \mathcal{L}}{\partial \nu}\right)_T = -e^{\delta_T}\, \mathbb{E}\left[\nabla f(X_T) \Big| \mathcal{F}_T\right] \end{cases} , \tag{4.1}$$

where we define the processes

$$\left(\frac{\partial \mathcal{L}}{\partial X}\right)_t = e^{\gamma_t + \alpha_t}\Big(\nabla h(X_t^{\nu^*} + e^{-\alpha_t}\nu_t^*) - \nabla h(X_t^{\nu^*}) - e^{-\alpha_t}\nabla^2 h(X_t^{\nu^*})\nu_t^* - e^{\beta_t}\nabla f(X_t^{\nu^*})\Big) \tag{4.2}$$

$$\left(\frac{\partial \mathcal{L}}{\partial \nu}\right)_t = e^{\gamma_t}\left(\nabla h(X_t^{\nu^*} + e^{-\alpha_t}\nu_t^*) - \nabla h(X_t^{\nu^*})\right) \ , \tag{4.3}$$

and where the process $\mathcal{M} = (\mathcal{M}_t)_{t\in[0,T]}$ is an $\mathcal{F}$-adapted martingale. As a consequence, if the solution to this FBSDE is unique, then it is the unique critical point of the functional $\mathcal{J}$ up to null sets.

Proof. See Appendix A □

Theorem 4.1 presents an analogue of the Euler-Lagrange equation with free terminal boundary. Rather than obtaining an ODE as in the classical result, we obtain an FBSDE, with backwards process $\left(\frac{\partial \mathcal{L}}{\partial \nu}\right)_t$, and the forward state processes consisting of of $\mathbb{E}\left[\left(\frac{\partial \mathcal{L}}{\partial X}\right)_t | \mathcal{F}_t\right]$, $\int_0^t \|\nu_u\|\, du$ and $X_t^{\nu^*}$.

For background on Forward-Backward Stochastic Differential Equations, we point reades to the papers [4, 17, 20], which formally define these equations and their solutions in full detail. Briefly, a solution to an FBSDE of the form (4.1) consists of a pair of processes $\left(\frac{\partial \mathcal{L}}{\partial \nu}, \mathcal{M}\right)$, which both satisfy the dynamics and the boundary condition of equation (4.1). Intuitively, the martingale part of the solution can be interpreted as a random process which guides $\left(\frac{\partial \mathcal{L}}{\partial X}\right)_t$ towards the terminal boundary at time T. We also note that solutions to (4.1) are non-Markovian in general, and should be regarded simply as maps from the filtration $\mathcal{F}_t$ into $\mathbb{R}^d$. Under the assumption that all state processes in this problem are jointly Markov, is possible to write the system (4.1) into an equivalent system of semi-linear PDEs, as is often done in the BSDE literature. Since we do not make this additional assumption, and since this often leads to additional complications, we will avoid taking this route.

Another important feature of equation (4.1), is that optimality relies on the projection of $\left(\frac{\partial \mathcal{L}}{\partial X}\right)_t$ onto $\mathcal{F}_t$. In fact, the only dependence of the optimizer dynamics on the random loss surface is through the projection of its gradients along the optimizers path onto $\mathcal{F}_t$, $\mathbb{E}\left[\nabla f(X_t)|\mathcal{F}_t\right]$ as is seen in the expression (4.2). Thus, the optimization algorithm makes use of past noisy gradient observations in order to make local gradient predictions. Local gradient predictions are updated in a Bayesian manner, where the a-priori model for ∇f is updated with path information contained in $\mathcal{F}_t$.

Theorem 4.1 provides us with a means of selecting amongst all admissible optimization algorithms $\mathcal{A}$ by choosing one that is optimal with respect to the expected action performance metric $\mathcal{J}$. Implicitly, these algorithms embed a notion of acceleration based on the kinetic energy induced by the Bregman divergence. To see this, notice that setting $h(x) = \frac{1}{2}\|x\|^2$ in equation (4.1) and letting f be non-random recovers the classical Newtonian equations of motion for a particle with unit mass. Tuning h in this context allows us to change the way momentum is accumulated by the optimizer as it descends the latent loss function.

Since $\mathcal{J}$ is not necessarily convex in general, it is not possible to determine whether solutions to the FBSDE (4.1) are either minima or maxima of the functional $\mathcal{J}$. We point out, however, that this also the case in general for both classical Lagrangian mechanics and in the treatment of the deterministic version of the problem in Wibisono et al. [28]. Nonetheless, we narrow our focus to optimizers which are solutions to the optimality FBSDE (4.1).

4.1. **Expected Rates of Convergence of the Continuous Algorithm.** Using the dynamics (4.1) we obtain a bound on the rate of convergence of the continuous optimization algorithm that is analogous to [28, Theorem 2.1]. By making additional assumptions on h, f and the hyperparameters α, γ and β, we can give upper bounds for the rate of convergence of the continuous algorithm. We proceed similarly to Wibisono et al. [28] and introduce the Lyapunov energy functional

$$\mathcal{E}_t = D_h(x^\star, X_t^{\nu^*} + e^{-\alpha_t}\nu_t) + e^{\beta_t}\left(f(X_t^{\nu^*}) - f(x^\star)\right) - [\nabla h(X^{\nu^*} + e^{-\alpha_t}\nu), X^{\nu^*} + e^{-\alpha_t}\nu]_t \,, \tag{4.4}$$

where we define $x^\star$ to be a global minimum of f. By showing that this quantity is a super-martingale with respect to the filtration $\mathcal{F}$, we obtain an upper bound for the expected rate of convergence from X_t towards the minimum.

Theorem 4.2 (Convergence Rate)**.** *Assume that the function f is almost surely convex and that the ideal scaling conditions* (4.5) *hold. Moreover assume that in addition to h having L-Lipschitz smooth gradients, h is also μ-strongly-convex with $\mu > 0$, and that the scaling conditions*

$$\dot{\gamma}_t = e^{\alpha_t} \quad \textit{and} \quad \dot{\beta}_t \leq e^{\alpha_t} \tag{4.5}$$

hold. Define $x^\star = \arg\min_{x\in\mathbb{R}^d} f(x)$ to be a global minimum of f. If $x^\star$ exists almost surely, the optimizer defined by FBSDE (4.1) *satisfies*

$$\mathbb{E}\left[f(X_t) - f(x^\star)\right] = O\left(e^{-\beta_t}\max\left\{1, \mathbb{E}\left[\,[e^{-\gamma_t}\mathcal{M}]_t\right]\right\}\right) \,, \tag{4.6}$$

where $[e^{-\gamma_t}\mathcal{M}]_t$ represents the quadratic variation of the process $e^{-\gamma_t}\mathcal{M}_t$, where $\mathcal{M}$ is the martingale defined in Theorem 4.1.

Proof. The proof is found in Appendix B. □

Theorem 4.2 shows that the rate of convergence of the continuous algorithm towards the optimum $x^\star$ of f is dominated by the two terms $e^{-\beta_t}$, which shrinks towards zero at a rate controlled by the growth of β, and the expected quadratic variation $\mathbb{E}\,[\,[e^{-\gamma}\mathcal{M}]_t]$. The latter term may increase or decrease in time depending on the exact model specification.

We can interpret the term $\mathbb{E}\,[\,[e^{-\gamma_t}\mathcal{M}]_t]$ as a penalty on the rate of convergence, which scales with the amount of noise present in our gradient observations as well as the algorithm hyperparamters. To see this, note that if there is no noise in our gradient observations, we obtain that $\mathcal{F}_t = \mathcal{G}_t$, and hence $\mathcal{M}_t \equiv 0$, recovering the exact rate of convergence $O(e^{-\beta_t})$ as in the deterministic case of [28]. If the noise in our gradient estimates is immense, we can expect $\mathbb{E}\,[\,[e^{-\gamma}\mathcal{M}]_t]$ to grow at an extremely fast rate and to counteract the shrinking effects of $e^{-\beta_t}$. We point out, however, that there will be a nontrivial dependence of $\mathbb{E}\,[\,[e^{-\gamma}\mathcal{M}]_t]$ on all model hyperparameters and the specific definition of the random variable f.

Remark 4.3. We do not assume that the conditions of Theorem 4.2 carry throughout the remainder of the paper. In particular, Sections 5 and 6 study models which may not guarantee almost-sure convexity of the latent loss function.

4.2. **Momentum-Based Representation of the Optimizer Dynamics.** Using a simple change of variables we may represent the dynamics of the FBSDE (4.1) in a simpler fashion, which will aid us in obtaining solutions to this system of equations. Let us define the *momentum process* $p = (p_t)_{t\in[0,T]}$ as

$$p_t = \left(\frac{\partial \mathcal{L}}{\partial \nu}\right)_t = e^{\gamma_t}\left(\nabla h(X_t^{\nu^*} + e^{-\alpha_t}\nu^*) - \nabla h(X_t^{\nu^*})\right) . \tag{4.7}$$

Noting that since h is convex, we have the property that $\nabla h^*(x) = (\nabla h)^{-1}(x)$, we may use equation (4.7) to write ν^* in terms of the momentum process as

$$\nu^* = e^{-\alpha_t}\left(\nabla h^*\big(\nabla h(X_t) + e^{-\gamma_t}p_t\big) - X_t\right) . \tag{4.8}$$

The introduction of this process allows us to represent the solution to the optimality FBSDE (4.1), and by extension the optimizer, in a much more tractable way. Re-writing the FBSDE (4.1) in terms of p_t, we find that

$$\begin{cases} dp_t = -\left\{e^{\gamma_t+\alpha_t+\beta_t}\mathbb{E}\left[\nabla f(X_t^{\nu^*})\big|\mathcal{F}_t\right] + \left(e^{\gamma_t}\nabla^2 h(X_t)\,\nu_t^* - e^{\alpha_t}p_t\right)\right\}dt + d\mathcal{M}_t \\ p_T = -e^{\delta_T}\mathbb{E}\left[\nabla f(X_T^{\nu^*})\big|\mathcal{F}_T\right] \end{cases} \tag{4.9}$$

where the dynamics of the forward process X^{ν^*} can be expressed as

$$dX_t^{\nu^*} = e^{\alpha_t}\left(\nabla h^*\Big(\nabla h(X_t^{\nu^*}) + e^{-\gamma_t}p_t\Big) - X_t^{\nu^*}\right)dt . \tag{4.10}$$

This particular change of variables corresponds exactly to the Hamiltonian representation of the optimizer's dynamics, which we show in Appendix C.

Writing out the explicit solution to the FBSDE (4.9), we obtain a representation for the optimizer's dynamics as

$$p_t = \mathbb{E}\left[\int_t^T e^{\gamma_u}\left\{e^{\alpha_u+\beta_u}\nabla f(X_u^{\nu^*}) + \left(\nabla^2 h(X_u)\,\nu_u^* - e^{\alpha_u-\gamma_u}p_u\right)\right\}du \; - e^{\delta_T}\nabla f(X_T^{\nu^*})\;\Big|\mathcal{F}_t\right] , \tag{4.11}$$

showing that optimizer's momentum can be represented as a time-weighted average of the expected future gradients over the remainder of the optimization and the term $e^{\gamma_t}\nabla^2 h(X_t)\,\nu_t^* - e^{\alpha_t}p_t$, where the weights are determined by the choice of hyperparameters α, β and γ. Noting that

$$\nabla^2 h(X_t)\,\nu_t^* - e^{\alpha_t-\gamma_t}p_t = \nabla^2 h(X_t)\nu_t^* - \left(\frac{\nabla h(X_t + e^{-\alpha_t}\nu_t^*) - \nabla h(X_t)}{e^{-\alpha_t}}\right) , \tag{4.12}$$

we find that the additional correction term in (4.11) can be interpreted as the remainder in the first-order Taylor expansion of the term $\nabla h(X_t + e^{-\alpha_t}\nu^*)$.

The representation (4.11) demonstrates optimizer does not only depend on the instantaneous value of gradients at the point $X_t^{\nu^*}$. Rather, we find that the algorithm's behaviour depends on the expected value of all future gradients that will be encountered over the remainder of the optimization process, projected onto the set of accumulated gradient information, $\mathcal{F}_t$. This is in stark contrast to most known stochastic optimization algorithms which only make explicit use of local gradient information in order to bring the optimizer towards an optimum.

4.3. **Zero-Order Singular Perturbation Approximation.** When h does not take the quadratic form $h(x) = \frac{1}{2}x^\intercal Mx$ for some positive-definite matrix M, the nonlinear dynamics of the FBSDE (4.1) or in the equivalent momentum form (4.9) make it difficult to provive a solution for general h. More precisely, the Taylor expansion term (4.12) constitutes the main obstacle in obtaining solutions in general.

In cases where the scaling parameter α_t is sufficiently large, we can assume that the Taylor expansion remainder term of equation (4.12) will become negligibly small. Hence, we may approximate the optimality dynamics of the FBSDE (4.9) by setting this term to zero. This can be interpreted as the zero-order term in a singular perturbation expansion of the solution to the momentum FBSDE (4.9).

Under the assumption that the Taylor remainder term vanishes, we obtain the approximation $\tilde{p}^{(0)} = (\tilde{p}^{(0)})_{t\in[0,T]}$ for the momentum, which we present in the following proposition.

Proposition 4.4 (Zero-Order Singular Perturbation (ZOSP))**.** *The linear FBSDE*

$$\begin{cases} d\tilde{p}_t^{(0)} = -e^{\alpha_t}\left\{e^{\beta_t}\,\mathbb{E}\left[\nabla f\left(X_t\right)|\mathcal{F}_t\right] + \tilde{p}_t^{(0)}\right\}\,dt + d\tilde{\mathcal{M}}_t^{(1)} \\ \tilde{p}_T^{(0)} = -e^{\delta_T}\mathbb{E}\left[\nabla f(X_T^{\nu^*})\middle|\mathcal{F}_T\right] \end{cases}, \tag{4.13}$$

admits a solution that can be expressed as

$$\tilde{p}_t^{(0)} = \mathbb{E}\left[\int_t^T e^{\gamma_t+\alpha_t+\beta_t}\,\nabla f\left(X_u\right)\,du - e^{\delta_T-\gamma_t}\nabla f(X_T^{\nu^*})\middle|\mathcal{F}_t\right], \tag{4.14}$$

provided that $\mathbb{E}\left[\int_0^T e^{\gamma_t+\alpha_t+\beta_t}\,\|\nabla f\left(X_u\right)\|\,du\right] < \infty$.

Although a general, model independent bound for the accuracy of such approximations is beyond the scope of this paper, it can still serve as a reasonable and computationally cheap alternative to attempting to solve the original problem dynamics directly with a BSDE numerical scheme. For more information on singular perturbation methods in the context of FBSDEs, see [13].

5. Stochastic Gradient Descent and Stochastic Mirror Descent

By specifying an appropriate model for f, we can recover mini-batch SGD and mirror descent from the optimizer dynamics we have presented this far. In order to achieve this, we model the gradients along the path of X_t as $\nabla f(X_t) = M_t^f$, where M_t^f is an L^2 martingale. We model the source of noise in the gradient observations as coming from another identically distributed and independent copy of M_t^f, which we write as M_t^ε. We therefore write the dynamics of the noisy gradient process as

$$dg_t = dM_t^f + \sigma dM_t^\varepsilon\,, \tag{5.1}$$

where $\sigma > 0$ represents the relative scale of the gradient noise. We can think of g_t in this model as the approximate gradient samples obtained from mini-batch samples of a risk function and σM_t^ε as the error of the mini-batch approximation.

Using symmetry we obtain a trivial solution to the gradient filter as $\mathbb{E}[\nabla f(X_t)|\mathcal{F}_t] = (1+\sigma)^{-1}g_t$, implying that the best estimate of the gradient at the point X_t will be the most recent mini-batch sample observed, re-scaled by a constant. Next, applying Fubini's theorem and the martingale property of g_t to

ZOSP approximation (4.14), and using the scaling conditions, we obtain

$$\tilde{p}_t^{(0)} = g_t\,(1+\sigma)^{-1}\left(\int_t^T e^{\alpha_u+\beta_u+\gamma_u}\,du - e^{\delta_T}\right) \tag{5.2}$$

$$= -g_t\,(1+\sigma)^{-1}\left(\Phi_0 + \int_0^t e^{\alpha_u+\beta_u+\gamma_u}\,du\right)\,, \tag{5.3}$$

where we define the constant $\Phi_0 = e^{\delta_T} - \int_0^T e^{\alpha_u+\beta_u+\gamma_u}\,du$, which we can make as large as we wish by adjusting the value of δ_T.

We may therefore interpret the momentum process as $\tilde{p}_t^{(0)} = -g_t\tilde{\Phi}_t$, where $\tilde{\Phi}_t > 0$ is a deterministic scalar learning rate, which is entirely determined by the choice of scaling parameters α, β, γ and δ. Using this solution to the momentum process, we write the continuous dynamics of the optimizer as

$$dX_t = e^{\alpha_t}\left(\nabla h^*\left(\nabla h(X_t) - g_t e^{-\gamma_t}\tilde{\Phi}_t\right) - X_t^{\nu^*}\right)\,dt\,. \tag{5.4}$$

To obtain discrete algorithms, we consider a discretization of the process $X_t^{\nu^*}$ on the finite mesh $\mathcal{T} = \{t_0 = 0\,,\ t_{k+1} = t_k + e^{-\alpha_{t_k}} : k \in \mathbb{N}\}$. Using an explicit Euler scheme to discretize X^{ν^*} on $\mathcal{T}$, we obtain the update rule

$$X_{t_{k+1}} = \nabla h^*\left(\nabla h(X_{t_k}) - g_{t_k}\tilde{\Phi}_{t_k}\right)\,, \tag{5.5}$$

corresponding exactly to mirror descent with minibatch gradients and time varying learning rate. Moreover, setting $h(x) = \frac{1}{2}\|x\|^2$, we exactly recover mini-batch SGD

$$X_{t_k} - X_{t_k} = -\tilde{\Phi}_{t_k} e^{-\gamma_{t_k}}\,g_{t_k}\,, \tag{5.6}$$

with the corresponding learning rate given by $\tilde{\Phi}_{t_k}e^{-\gamma_{t_k}}$.

This derivation demonstrates that a prior which imposes minimal assumptions on the evolution of gradients of the optimizer will yield stochastic gradient descent. In particular, the gradient model (5.1) can be rougly interpreted as assuming that gradients behave as random walks over the path of the optimizer, moving with zero-mean independent increments. Moreover, the optimal gradient filter $\mathbb{E}[\nabla f(X_t)|\mathcal{F}_t] = (1+\sigma)^{-1}g_t$ shows that the noisy gradients should be scaled down as the scale of the noise σ in the g_t increases. Since in practice we find that the minibatch noise scales as $\sigma \approx \ell^{-\frac{1}{2}}$, where ℓ is the size of the minibatch, this observation suggests that minibatch gradient should be scaled down by a factor of $(1+\ell^{-\frac{1}{2}})^{-1}$.

6. Kalman Gradient Descent and Momentum Methods

Using diffusive models for gradients, we can recover both the Kalman Gradient Descent algorithm from [27] and the classical momentum-based optimization methods of [21]. In order to do this, we must first introduce linear state-space models for the time evolution of gradients. In particular, [11, 14, 24, 25] show an equivalence between these models and time-homogeneous Gaussian Processes (GP).

6.1. State-Space Models for Gradients.

Let us assume that each component of $\nabla f(X_t) = (\nabla_i f(X_t))_{i=1}^d$ is modelled independently as a linear diffusive process. Specifically, we assume that there exist processes $y_i = (y_{i,t})_{t\geq 0}$ so that for each i, $y_{i,t} \in \mathbb{R}^m$ and $y_{i,t}$ is the solution to the linear SDE

$$dy_{i,t} = -Ay_{i,t}dt + LdW_{i,t}\,, \tag{6.1}$$

where $A, L \in \mathbb{R}^{m\times m}$ are both positive definite matrices and each of the $W_i = (W_{i,t})_{t\geq 0}$ are independent m-dimensional Brownian Motions.

Our state-space model therefore assumes that each element of the gradient can be written as

$$\nabla_i f(X_t) = b^\intercal y_{i,t} \tag{6.2}$$

where $b \in \mathbb{R}^d$. Next, we assume that we may write each element of the noisy gradient process as the solution to the SDE

$$g_{i,t} = b^\intercal y_{i,t} + \sigma dB_{i,t}\,, \tag{6.3}$$

where $\sigma > 0$ and where $B_i = (B_{i,t})_{t\geq 0}$ are independent Brownian Motions.

This model can be interpreted as placing a prior model on gradients so that each $\nabla_i f(X_t)$ is an independent Gaussian Process with a stationary kernel $k(t)$. Similarly, the noisy gradient process can then be informally interpreted as $\frac{dg_t}{dt} = \nabla_i f(X_t) + \sigma\xi_{i,t}$, where each of the ξ_i are i.i.d. space-time white noise processes. For more information on the connection between state space models and Gaussian Processes, see any of [11, 14, 24, 25].

Noting that under this model we have $\mathbb{E}[\nabla_i f(X_{t+h})|y_{i,t}] = b^\intercal e^{-Ah} y_{i,t}$, we find that this model assumes that gradients decrease in exponentially in magnitude as a function of time over the course of the optimization as a rate controlled by the matrix A. The parameters σ and $\mathcal{L}$ can be interpreted as controlling the scale of the variation in the noise and signal processes, respectively.

Using the state-space model, we obtain that the filter can be expressed as $\mathbb{E}[\nabla_i f(X_t)|\mathcal{F}_t] = b^\intercal \hat{y}_{i,t}$, where $\hat{y}_{i,t} = \mathbb{E}\left[y_{i,t}\middle|\mathcal{F}_t\right]$. The process $\hat{y}_{i,t}$ is expressed as the solution to the Kalman-Bucy[1] filtering equations

$$d\hat{y}_{i,t} = -A\hat{y}_{i,t}\,dt + \sigma^{-1}\hat{P}_t\, b d\hat{B}_{i,t} \tag{6.4}$$

$$\frac{d\hat{P}_t}{dt} = -A\hat{P}_t - \hat{P}_t^\intercal A - \sigma^{-2}\hat{P}_t^\intercal b\, b^\intercal \hat{P}_t + LL^\intercal\,, \tag{6.5}$$

with the initial conditions $\hat{y}_{i,0} = 0$ and $\hat{P}_0 = \mathbb{E}\left[y_{i,0}y_{i,0}^\intercal\right]$, and where we define innovations process $\hat{B}_{i,t} = \sigma^{-1}\left(g_{i,t} - b^\intercal \hat{y}_{i,t}\right)$ with the property that each $\hat{B}_i$ is an independent $\mathcal{F}$-adapted Brownian motion. Note that equation (6.5) constitutes a deterministic Riccati ODE, and whose solution does not vary across $i = 1, \ldots, d$.

Moreover, since the dynamics of $\hat{y}_{i,t}$ remain linear, we can easily compute the expected future value of each gradient as

$$\mathbb{E}\left[\nabla_i f(X^{\nu^*}_{t+h})|\mathcal{F}_t\right] = b^\intercal e^{-Ah}\hat{y}_{i,t}\,, \tag{6.6}$$

where e^{-Ah} is the matrix exponential.

6.2. **Kalman Gradient Descent.** We now turn to using this model within the optimization equations in order to derive discrete optimization algorithms. Using the state space gradient model, we can derive an exact solution to the ZOSP in Proposition (4.4).

[1]For more information on continuous time filtering and the Kalman-Bucy filter we refer the reader to the text of Bensoussan [3] or the lecture notes of Van Handel [26].

Proposition 6.1 (State-Space Model Solution to the ZOSP)**.** *Assume that the gradient state-space model described in Section 6.1 holds. The solution to the ZOSP momentum process can therefore be expressed as*

$$\tilde{p}_{i,t}^{(0)} = -b^\intercal \Phi_t \hat{y}_{t,i} \ , \tag{6.7}$$

where $\Phi : [0,T] \to \mathbb{R}^{m\times m}$ *is defined as*

$$\Phi_t = \Phi_0 + \int_0^t e^{\alpha_u + \beta_u + \gamma_u \, ds} e^{-A(u-t)} \, du \tag{6.8}$$

where e^A *represents the matrix exponential, and where*

$$\Phi_0 = e^{\delta_T} e^{AT} - \int_0^T e^{\alpha_u + \beta_u + \gamma_u} e^{-Au} \, du \tag{6.9}$$

can be chosen to have arbitrarily large eigenvalues by scaling δ_T*.*

Proof. The result is an immediate consequence of inserting (6.6) into equation (4.14). □

The solution to the momentum process is therefore represented in each coordinate as a linear function of the filter $\hat{y}_{i,t}$, scaled by a matrix-valued learning rate Φ_t. In particular, the form of the solution closely resembles the form obtained in Section 5, where we replace g_t by a gradient filter over past gradients instead.

In order to recover Kalman Gradient Descent, we must both discretize the process $X_t^{\nu^*}$ and each filter process $\hat{y}_i$. Discretizing equations (6.1) and (6.3) over the mesh $\mathcal{T} = \{t_0 = 0 \, , \ t_{k+1} = t_k + e^{-\alpha_{t_k}} : k \in \mathbb{N}\}$ with an Euler-Maruyama scheme, we obtain the discrete evolution equations

$$y_{i,t_k} \approx (I - e^{-\alpha_{t_k}} A) y_{i,t_k} + L e^{-\alpha_t} w_{i,k} \tag{6.10}$$

$$g_{i,t_k} \approx b^\intercal y_{i,t_k} + e^{-\alpha_t} \xi_{i,k} \ , \tag{6.11}$$

where each of the $\xi_{i,k}$ and $w_{i,k}$ are standard Gaussian random variables of appropriate size. The filter $\hat{y}_{i,k} = \mathbb{E}\left[y_{i,t_k} | \sigma(g_{i,t_{k'}})_{k'=1}^k\right]$ for the discrete equations can be written as the solution to the Kalman filtering equation. To obtain the exact discrete filter dynamics see [11].

Next, discretizing the process X^{ν^*} over $\mathcal{T}$ with an Euler scheme, just as in Section 5, we obtain discrete dynamics for the optimizer, as

$$X_{t_{k+1}} = \nabla h^* \left(\nabla h(X_{t_k}) - \left(b^\intercal e^{-\gamma_{t_k}} \Phi_{t_k} \hat{y}_{i,t_k} \right)_{i=1}^d \right) \ , \tag{6.12}$$

yielding a generalized version of Kalman gradient descent with deterministic learning rate $e^{-\gamma_{t_k}} \Phi_{t_k}$ and with mirror descent steps. Equation (6.12) generalizes on the Kalman gradient descent algorithm of Vuckovic [27] by combining elements of mirror descent, as well as multiple states for each element of the gradient. Setting $h(x) = \frac{1}{2}\|x\|^2$, $m = 1$ and $b = 1$ recovers the original algorithm found in [27] with a time-varying learning rate.

6.3. **Momentum and Aggregated Momentum Methods.** We now demonstrate how momentum methods can be represented as a special asymptotic case of Kalman gradient descent. If we consider the case where $t \to \infty$ in equation (6.5), we find that $\hat{P}_t \to \hat{P}_\infty \in \mathbb{R}^{m\times m}$, where $\hat{P}_\infty$ is the solution to the continuous algebraic Riccati equation

$$0 = -A\hat{P}_\infty - \hat{P}_\infty^\intercal A - \sigma^{-2} \hat{P}_\infty^\intercal b \, b^\intercal \hat{P}_\infty + LL^\intercal \ , \tag{6.13}$$

which is guaranteed to have a solution under our model assumptions. In this limiting case, the filtering equations (6.4) can be represented as

$$d\hat{y}_{i,t} = -A\hat{y}_{i,t}\,dt + \sigma^{-1}\hat{P}_\infty\, b^\intercal d\hat{B}_{i,t}\ . \tag{6.14}$$

Now, proceeding with an Euler-Maruyama discretization of this process over $\mathcal{T}$, and using the definition of $\hat{B}_{i,t}$ we obtain the discrete evolution dynamics

$$\hat{y}_{i,t_{k+1}} \approx \left(I - \tilde{A}_k\right)\hat{y}_{i,t_k} + \tilde{L}_k\,\Delta_k g_i \tag{6.15}$$

where $\tilde{A}_k = Ae^{-\alpha_{t_k}} - \sigma^{-2}e^{-\alpha_t}\hat{P}_\infty\, b\, b^\intercal$, $\tilde{L}_k = \sigma^{-2}e^{-\alpha_t}\hat{P}_\infty\, b\, b^\intercal$ and $\Delta_k g_i = g_{i,t_{k+1}} - g_{i,t_k}$. Notice, however that when we use an Euler-Maruyama approximation to $\boldsymbol{g}_t$, we obtain

$$e^{\alpha_t}\Delta_k g \approx \nabla f(X_{t_k}) + \zeta_{i,k}\ , \tag{6.16}$$

where $\zeta_{i,k}$ are independent standard Gaussian random variables.

Equation (6.16) therefore demonstrates that each of the $\Delta_k g$ terms correspond to a re-scaled noisy observation of the gradient. Moreover, we find from equation (6.15) that the asymptotic discretized Kalman filter dynamics correspond to a generalized momentum rule where we keep track of m state processes which evolve via vector valued equations.

This generalized algorithm can be considered most similar to the aggregated momentum algorithm of [16], since it keeps track of m separate momentum states which each evolve according to their own momentum rate as given by the parameters $\tilde{A}_k$ and $\tilde{L}_k$. The main difference between the two is that this version of the algorithm allows for interaction between the m momentum states in their evolution equations. Lastly, using the very same discretization of X^{ν^*} as in Section 6.2, we obtain the update rule

$$X_{t_{k+1}} = \nabla h^*\left(\nabla h(X_{t_k}) - \left(b^\intercal e^{-\gamma_{t_k}}\Phi_{t_k}\hat{y}_{i,t_k}\right)_{i=1}^d\right) \tag{6.17}$$

$$\hat{y}_{i,t_{k+1}} = \left(I - \tilde{A}\right)\hat{y}_{i,t_k} + \tilde{L}\,\Delta_k g_i\ . \tag{6.18}$$

In the particular case where $m = 1$, $b = 1$, and where $h(x) = \frac{1}{2}\|x\|^2$, we recover the classical momentum rule by setting the hyper-parameter $e^{-\alpha_t} = \epsilon$. Under this assumption, we find that both $\tilde{A}$ and $\tilde{L}$ are constants which do not vary with k, and hence by setting $\tilde{L} = \tilde{A} \in (0, 1)$ we obtain

$$X_{t_{k+1}} = -e^{-\gamma_{t_k}}\Phi_{t_k}\hat{y}_{t_k} \tag{6.19}$$

$$\hat{y}_{i,t_{k+1}} = \left(1 - \tilde{A}\right)\hat{y}_{i,t_k} + \tilde{A}\,\Delta_k g_i\ , \tag{6.20}$$

which is the classical momentum rule with weight $\tilde{A}$ on each new gradient observation and decay rate $1 - \tilde{A}$.

Thus, we find that momentum methods can be interpreted as a special case of applying a linear SDE prior to gradients, which depend strictly on time. In light of the state space model in Section 6.1, we find that momentum methods implicitly assume that gradients decay exponentially in time, with the exact rate depending on the chosen model parameters. This also suggests that the weighting parameters in a momentum scheme should be chosen as a function of (i) how fast gradients are expected to decay on average over the course of the optimization (ii) how much noise we expect in the evolution of gradients and (iii) the scale of the observational noise in the minibatch gradient estimates.

7. Discussion and Future Work

Over the course of the paper we present a variational framework on optimizers, which interprets the task of stochastic optimization as an inference problem on a latent surface that we wish to optimize. By solving a variational problem over continuous optimizers with asymmetric information, we find that optimal algorithms should satisfy a system of FBSDEs projected onto the filtration $\mathcal{F}$ generated by the noisy observations of the latent process.

By solving these FBSDEs and obtaining continuous-time optimizers, we find a direct relationship between the measure assigned to the latent surface and its relationship to how data is observed. In particular, assigning simple prior models to the pair of processes $(\nabla f(X_t), g_t)_{t\in[0,T]}$, recovers a number of well known and widely used optimization algorithms. The fact that this framework can recover all of these algortihms in a natural way begs further study. In particular, it is still an open question whether it is possible to recover other stochastic algorithms via this framework, particularily those with second-order scaling adjustments such as ADAM or AdaGrad.

From a more technical perspective, the intent is to further explore properties of the optimization model presented here and the form of the algorithms it suggests. In particular, the optimality FBSDE 4.1 is nonlinear, high-dimensional and tractable in general, making it difficult to use existing FBSDE approximation techniques, so new tools may need to be developed to understand the full extent of its behaviour.

Lastly, numerical work on the algorithms generated by this framework would give some insights as to which prior gradient models work well when discretized. The extension of simplectic and quasi-simplectic stochastic integrators applied to the BSDEs and SDEs that appear in this paper also has the potential for interesting future work.

Appendix A. Proofs relating to Theorem 4.1

Before going forward with the main part of the proof, we first present a lemma for the computation of the Gâteaux derivative of $\mathcal{J}$.

Lemma A.1. *The functional $\mathcal{J}$ is everywhere Gâteaux differentiable in $\mathcal{A}$. The Gâteaux at a point $\nu \in \mathcal{A}$ in the direction $\tilde{\omega} = \omega - \nu$ for $\omega \in \mathcal{A}$ takes the form*

$$\langle D\mathcal{J}(\nu), \tilde{\omega} \rangle = \mathbb{E}\left[\int_0^T \left\langle \omega_t \,,\ \frac{\partial \mathcal{L}\,(t, X_t^\nu, \nu_t)}{\partial \nu} - \mathbb{E}\left[\int_t^T \frac{\partial \mathcal{L}\,(u, X_u^\nu, \nu_u)}{\partial X}\, du - e^{\delta_T}\, \nabla f\,(X_T^\nu) \,\middle|\, \mathcal{F}_t \right] \right\rangle dt \right] . \tag{A.1}$$

Proof. If we assume that the conditions of Leibniz' rule hold, we may compute the Gâteax derivative as

$$\begin{aligned}
\partial_\rho \mathcal{J}\,(\nu + \rho\,\tilde{\omega}) &= \partial_\rho \mathbb{E}\left[\int_0^T \mathcal{L}\left(t, X_t^{\nu+\rho\,\tilde{\omega}}, \nu_t + \rho\,\tilde{\omega}_t\right) dt + e^{\delta_T}\left(f(X_T^{\nu+\rho\,\tilde{\omega}}) - f(x^\star) \right) \right] \\
&= \mathbb{E}\left[\int_0^T \partial_\rho \mathcal{L}\left(t, X_t^{\nu+\rho\,\tilde{\omega}}, \nu_t + \rho\,\tilde{\omega}_t\right) dt + e^{\delta_T} \partial_\rho f(X_T^{\nu+\rho\,\tilde{\omega}}) \right] \\
&= \mathbb{E}\left[\int_0^T \left\{ \left\langle \frac{\partial \mathcal{L}\,(t, X_t^\nu, \nu_t)}{\partial X} \,,\ \int_0^t \tilde{\omega}_u\, du \right\rangle + \left\langle \frac{\partial \mathcal{L}\,(t, X_t^\nu, \nu_t)}{\partial \nu} \,,\ \tilde{\omega}_t \right\rangle \right\} dt + \left\langle \int_0^T \tilde{\omega}_u\, du, \Phi \nabla f(X_T^\nu) \right\rangle \right] ,
\end{aligned} \tag{A.2}$$

where we have

$$\frac{\partial \mathcal{L}\,(t, X, \nu)}{\partial X} = e^{\gamma_t + \alpha_t}\Big(\nabla h(X + e^{-\alpha_t}\nu) - \nabla h(X) - e^{-\alpha_t}\nabla^2 h(X)\nu - e^{\beta_t}\nabla f(X) \Big) \tag{A.3}$$

$$\frac{\partial \mathcal{L}\,(t, X, \nu)}{\partial \nu} = e^{\gamma_t}\left(\nabla h(X + e^{-\alpha_t}\nu) - \nabla h(X) \right) . \tag{A.4}$$

Note here that the derivative in f is path-wise for every fixed realization of the function f. Since $f \in C^1$, we have that ∇f is also well-defined for every realization of f.

To ensure that this computation is valid, and that the conditions of the Leibniz rule are met, due to the continuity of (A.2) in $\tilde{\omega}$, is sufficient for us to show that the integrals in equation (A.2) are bounded for any $\tilde{\omega}$ and ν. First, note that by the Young and Jensen inequalities,

$$\mathbb{E}\left[\left\langle \int_0^T \tilde{\omega}_u\, du, \Phi \nabla f(X_T^\nu) \right\rangle \right] \le \frac{1}{2}\mathbb{E}\left[\int_0^T \|\tilde{\omega}_u\|^2\, du + \Phi \|\nabla f(X_T^\nu)\|^2 \right] < \infty , \tag{A.5}$$

where the boundedness holds from the fact that $\tilde{\omega} \in \mathcal{A}$ and that $\mathbb{E}\|f(x)\|^2 < \infty$ for all $x \in \mathbb{R}^d$.

Next, we focus on the left part of equation (A.2). By the Cauchy-Schwarz and Young inequalities, we have

$$\left| \left\langle \frac{\partial \mathcal{L}\,(t, X_t^\nu, \nu_t)}{\partial X} \,,\ \int_0^t \tilde{\omega}_u\, du \right\rangle + \left\langle \frac{\partial \mathcal{L}\,(t, X_t^\nu, \nu_t)}{\partial \nu} \,,\ \tilde{\omega}_t \right\rangle \right| \le \left\| \frac{\partial \mathcal{L}\,(t, X_t^\nu, \nu_t)}{\partial X} \right\| \left\| \int_0^t \tilde{\omega}_u\, du \right\| + \left\| \frac{\partial \mathcal{L}\,(t, X_t^\nu, \nu_t)}{\partial \nu} \right\| \|\tilde{\omega}_t\| \tag{A.6}$$

$$\le \frac{1}{2}\left\{ \left\| \int_0^t \tilde{\omega}_u\, du \right\|^2 + \left\| \frac{\partial \mathcal{L}\,(t, X_t^\nu, \nu_t)}{\partial X} \right\|^2 + \| \frac{\partial \mathcal{L}\,(t, X_t^\nu, \nu_t)}{\partial \nu} \|^2 + \|\tilde{\omega}_t\|^2 \right\} . \tag{A.7}$$

Using the L-Lipschitz property of the gradients of h, we can also bound the partial derivatives of the Lagrangian with the triangle inequality as

$$\begin{aligned}
\left\| \frac{\partial \mathcal{L}\,(t, X_t^\nu, \nu_t)}{\partial X} \right\| &\le e^{\gamma_t + \alpha_t} \left\| \nabla h(X + e^{-\alpha_t}\nu) - \nabla h(X) \right\| + e^{\gamma_t} \|\nabla^2 h(X)\nu\| + e^{\beta_t + \gamma_t + \alpha_t} \|\nabla f(X)\| \\
&\le L(e^{\gamma_t + \alpha_t} + e^{\gamma_t})\|\nu\| + e^{\beta_t + \gamma_t + \alpha_t} \|f(X)\| \\
&\le C_0\left(\|\nu\| + \|\nabla f(X)\| \right) \\
\left\| \frac{\partial \mathcal{L}\,(t, X, \nu)}{\partial \nu} \right\| &\le e^{\gamma_t} \left\| \nabla h(X + e^{-\alpha_t}\nu) - \nabla h(X) \right\| \\
&\le e^{\gamma_t} L\, \|\nu\| \\
&\le C_0\ \|\nu\| \ ,
\end{aligned}$$

where $C_0 = \sup_{t\in[0,T]}\{e^{\alpha_t + \gamma_t} + e^{\gamma_t} + e^{\alpha_t + \gamma_t + \beta_t}\}$ is bounded by the assumption that α, β, γ are continuous in $[0, T]$.

Using the above result, and applying Young's inequality to the previous result, we can upper bound equation (A.7) as

$$\text{(A.7)} \leq 32\,(1+C)\,\left\{1+\int_0^T \|\tilde{\omega}_u\|^2\,du+\|\nu_t\|^2+\|\tilde{\omega}_t\|^2+\|\nabla f\,(X_t)\,\|^2\right\} \tag{A.8}$$

$$\leq 64\,(1+C)\,\left\{1+\int_0^T \|\omega_u\|^2\,du+\int_0^T \|\nu_u\|^2\,du+\|\nu_t\|^2+\|\omega_t\|^2+\|\nabla f\,(X_t)\,\|^2\right\}\,, \tag{A.9}$$

where the number 32 is chosen to be much larger than what is strictly necessary by Young's inequality. Notice here that by the definition of $\mathcal{A}$, this forms an integrable upper bound to the left integral of equation (A.2), validating our use of Leibniz's rule, and showing that $\mathcal{J}$ is indeed Gâteaux integrable.

Now that integrability concerns have been dealt with, we can proceed with the computation of the Gâteaux derivative. By applying integration by parts to the left side of equation (A.10) and moving the right hand side into the integral, we obtain

$$\partial_\rho \mathcal{J}\,(\nu+\rho\,\tilde{\omega}) = \mathbb{E}\left[\int_0^T \left\langle \tilde{\omega}_t\,,\ \frac{\partial \mathcal{L}\,(t,X_t^\nu,\nu_t)}{\partial \nu}-\int_t^T \frac{\partial \mathcal{L}\,(u,X_u^\nu,\nu_u)}{\partial X}\,du-e^{\delta_T}\nabla f(X_T^\nu)\right\rangle\,dt\right]$$

Using the tower property and Fubini's theorem on the right, we get

$$\langle D\mathcal{J}(\nu),\tilde{\omega}\rangle = \mathbb{E}\left[\int_0^T \left\langle \tilde{\omega}_t\,,\ \frac{\partial \mathcal{L}\,(t,X_t^\nu,\nu_t)}{\partial \nu}-\mathbb{E}\left[\int_t^T \frac{\partial \mathcal{L}\,(u,X_u^\nu,\nu_u)}{\partial X}\,du+e^{\delta_T}\nabla f(X_T^\nu)\,\middle|\mathcal{F}_t\right]\right\rangle\,dt\right]\,, \tag{A.10}$$

as desired. □

A.1. **Proof of Theorem 4.1.** Using the representation of the Gâteux derivative of $\mathcal{J}$ brought forth by Lemma A.1, we may proceed with the proof of Theorem 4.1.

Proof of Theorem 4.1. The goal is to show that the BSDE (4.1) is a necessary and sufficient condition for ν^* to be a critical point of $\mathcal{J}$. For any Gâteaux differentiable function $\mathcal{J}$, a necessary and sufficient condition for a point $\nu^* \in \mathcal{A}$ to be a critical point is that its Gâteaux derivative vanished in any valid direction. Lemma A.1 shows that the Gâteaux derivative takes the form of equation (A.1). Therefore, all that remains is to show that the FBSDE 4.1 is a necessary and sufficient condition for equation (A.1) to vanish.

Sufficiency. We will show that equation (A.1) vanishes when the FBSDE (4.1) holds. Assume that there exists a solution to the FBSDE (4.1) satisfying $\nu^* \in \mathcal{A}$. We may then express the solution to the FBSDE explicitly as

$$\left(\frac{\partial \mathcal{L}}{\partial \nu}\right)_t = \mathbb{E}\left[\int_t^T \left(\frac{\partial \mathcal{L}}{\partial X}\right)_u\,du\,-e^{\delta_T}\nabla f(X_T^\nu)\,\middle|\mathcal{F}_t\right]\,.$$

Inserting this into the right side of (A.1), we find that $\langle D\mathcal{J}(\nu),\omega\rangle$ vanishes for all $\omega \in \mathcal{A}$, demonstrating sufficiency.

Necessity. Conversely, let us assume that $\langle D\mathcal{J}(\nu),\omega-\nu\rangle = 0$ for all $\omega \in \mathcal{A}$ and for some $\nu \in \mathcal{A}$ for which the FBSDE (4.1) is not satisfied. We will show by contradiction that this statement cannot hold by choosing a direction in which the Gâteax derivative does not vanish. Consider the choice

$$\omega_t^\rho = \nu_t+\rho\,\left(\frac{\partial \mathcal{L}\,(t,X_t^\nu,\nu_t)}{\partial \nu}-\mathbb{E}\left[\int_t^T \frac{\partial \mathcal{L}\,(u,X_u^\nu,\nu_u)}{\partial X}\,du-e^{\delta_T}\nabla f(X_T^\nu)\middle|\mathcal{F}_t\right]\right)\,, \tag{A.11}$$

for some sufficiently small $\rho > 0$. We will first show that $\omega^\rho \in \mathcal{A}$ for some $\rho > 0$.

First, note that clearly ω^ρ must be $\mathcal{F}_t$-adapted, and we have $\omega^0 = \nu_t$. Moreover, note that since $\nu \in \mathcal{A}$, we have that $\mathbb{E}\int_0^T \|\nu_t\|^2+\|\nabla f(X^\nu)\|^2\,dt < \infty$, that $\omega^0 = \nu$. Notice that by the continuity of ∇f and the definition of X, the expression

$$\mathbb{E}\int_0^T \|\omega_t^\rho\|^2+\|\nabla f(X^{\omega^\rho})\|^2\,dt \tag{A.12}$$

is continuous in ρ. Since (A.12) is bounded for $\rho = 0$, by continuity there exists some $\rho > 0$ for which (A.12) is bounded and by extension where $\omega^\rho \in \mathcal{A}$ for this same value of ρ.

Inserting (A.11) into the Gâteaux derivative (A.1), we get that

$$\langle D\mathcal{J}(\nu),\omega^\rho-\nu\rangle = \rho\,\mathbb{E}\left[\int_0^T \left\|\frac{\partial \mathcal{L}\,(t,X_t^\nu,\nu_t)}{\partial \nu}-\mathbb{E}\left[\int_t^T \frac{\partial \mathcal{L}\,(u,X_u^\nu,\nu_u)}{\partial X}\,du-e^{\delta_T}\nabla f(X_T^\nu)\,\middle|\mathcal{F}_t\right]\right\|^2\,dt\right]\,, \tag{A.13}$$

which is strictly positive unless the FBSDE (4.1) is satisfied, thus forming a contradiction and demonstrating that the condition is necessary. □

Appendix B. Proof of Theorem 4.2

Proof. The proof of this theorem is broken up into multiple parts. The idea will be to first show that the energy functional $\mathcal{E}$ is a super-martingale with respect to $\mathcal{F}_t$, and then to use this property to bound the expected distance to the optimum. Lastly, we bound a quadratic co-variation term which appears within these equations to obtain the final result.

Before delving into the proof, we introduce standard notation for semi-martingale calculus. We use the noation $dY_t = dY_t^c + \Delta Y_t$ to indicate the increments of the continuous part Y^c of a process Y and its discontinuities $\Delta Y_t = Y_t - Y_{t-}$, where we use the notation $t-$ to indicate the left limit of the process. We use the notation $[Y,Z]_t$ to represent the quadratic co-variation of two processes Y and Z. This quadratic variation term can be decomposed into $d[Y,Z]_t = d[Y,Z]_t^c + \langle \Delta Y_t, \Delta Z_t \rangle$, where $[Y,Z]_t^c$ represents the quadratic covariation between Y^c and Z^c, and where $\langle \Delta Y_t, \Delta Z_t \rangle$ represents the inner product of their discontinuities at t. For more information on semi-martingale calculus and the associated notation, see Jacod and Shiryaev [12, Sections 3-5].

Dynamics of the Bregman Divergence. The idea will now be to show that the energy functional $\mathcal{E}$, defined in equation (4.4), is a super-martingale with respect to the visible filtration $\mathcal{F}_t$.

Using Itô's formula and Itô's product rule for càdlàg semi-martingales [12][Theorem 4.57], as well as the short-hand notation $Y_t = X_t + e^{-\alpha_t}\nu_t^*$, we obtain

$$
\begin{aligned}
dD_h(x^\star, Y_t) &= -\left\{ \langle \nabla h(Y_t), dY_t^c \rangle + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 h(Y_t)}{\partial x_i \partial x_i} d\left[Y_i, Y_j\right]_t^c + \Delta h(Y_t) \right\} - \left\{ \langle d\nabla h(Y_t), x^\star - Y_t \rangle - \langle \nabla h(Y_t), dY_t \rangle - d\left[\nabla h(Y), Y\right]_t \right\} \\
&= -\left\{ \langle \nabla h(Y_t), -\Delta Y_t \rangle + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 h(Y_t)}{\partial x_i \partial x_i} d\left[Y_i, Y_j\right]_t^c + \Delta h(Y_t) \right\} - \left\{ \langle d\nabla h(Y_t), x^\star - Y_t \rangle - \sum_{i,j=1}^{d} \frac{\partial^2 h(Y_t)}{\partial x_i \partial x_i} d\left[Y_i, Y_j\right]_t^c - \langle \Delta\left(\nabla h(Y_t)\right), \Delta \right. \right. \\
&= -\left\{ \Delta h(Y_t) - \langle \nabla h(Y_t), \Delta Y_t \rangle \right\} - \langle d\nabla h(Y_t), x^\star - Y_t \rangle + \left\{ \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 h(Y_t)}{\partial x_i \partial x_i} d\left[Y_i, Y_j\right]_t^c + \langle \Delta\left(\nabla h(Y_t)\right), \Delta Y_t \rangle \right\} ,
\end{aligned}
$$

where from line 1 to 2, we use the identity $d[\nabla g(Y), Y]_t = \sum_{i,j} \frac{\partial^2 g(Y_t)}{\partial x_i \partial x_j} d[Y_i, Y_j]_t^c + \langle \Delta(\nabla g(Y_t)), \Delta Y_t \rangle$ for any C^2 function g.

Note that since h is convex, $\nabla^2 h$ must have positive eigenvalues, and hence $\frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 h(Y_t)}{\partial x_i \partial x_i} d\left[Y_i, Y_j\right]_t^c \geq 0$. The convexity of h also implies that $\langle \nabla h(x) - \nabla h(y), x - y \rangle \leq 0$, and therefore we get $\langle \Delta\left(\nabla h(Y_t)\right), \Delta Y_t \rangle \geq 0$. The convexity of h also implies that $\Delta h(Y_t) - \langle \nabla h(Y_t), \Delta Y_t \rangle \geq 0$. Combining these observations, we find that

$$
dD_h(x^\star, Y_t) \leq -\langle d\nabla h(Y_t), x^\star - Y_t \rangle + \left\{ \sum_{i,j=1}^{d} \frac{\partial^2 h(Y_t)}{\partial x_i \partial x_i} d\left[Y_i, Y_j\right]_t^c + \langle \Delta\left(\nabla h(Y_t)\right), \Delta Y_t \rangle \right\} \tag{B.1}
$$

$$
= -\langle d\nabla h(Y_t), x^\star - Y_t \rangle + \left[\nabla h(Y), Y\right]_t . \tag{B.2}
$$

Super-martingale property of $\mathcal{E}$. Noting that under the scaling conditions (4.5) to the optimality FBSDE (4.1), we obtain the dynamics

$$
d\nabla h(X_t^{\nu^*} + e^{-\alpha_t}\nu^*) = -e^{\alpha_t + \beta_t} \mathbb{E}\left[\nabla f(X_t^{\nu^*}) \middle| \mathcal{F}_t \right] dt + d\tilde{\mathcal{M}}_t . \tag{B.3}
$$

Inserting this in to the dynamics of for the energy functional, and applying the upper bound (B.2), we find that

$$
d\mathcal{E}_t \leq -\langle d\nabla h(Y_t), x^\star - Y_t \rangle + \dot{\beta}_t e^{\beta_t}\left(f(X_t) - f(x^\star)\right) dt + e^{\beta_t}\langle \nabla f(X_t), \nu_t \rangle dt \tag{B.4}
$$

$$
= \left\langle e^{\alpha_t + \beta_t} \mathbb{E}[\nabla f\left(X_t\right) | \mathcal{F}_t] dt - d\mathcal{M}_t \ , \ x^\star - Y_t \right\rangle + \dot{\beta}_t e^{\beta_t}\left(f(X_t) - f(x^\star)\right) dt + e^{\beta_t}\langle \nabla f(X_t), \nu_t \rangle dt \tag{B.5}
$$

$$
= -\left\{ D_f(x^\star, Y_t) + \left(e^{\alpha_t} - \dot{\beta}_t\right) e^{\beta_t}\left(f(X_t) - f(x^\star)\right) \right\} dt + d\mathcal{M}_t' , \tag{B.6}
$$

where we use the notation $\mathcal{M}_t'$ to represent the $\mathcal{F}_t$-martingale defined as

$$
d\mathcal{M}_t' = \left\langle e^{\alpha_t + \beta_t}\left(\mathbb{E}[\nabla f\left(X_t\right) | \mathcal{F}_t] - f(X_t)\right) dt - d\mathcal{M}_t \ , \ x^\star - Y_t \right\rangle . \tag{B.7}
$$

Now note that due to the assumed convexity of f, we have that $D_f(x^\star, Y_t)$ is almost surely non-negative. Second, by the scaling conditions in equation (4.5), $e^{\alpha_t} - \dot{\beta}_t$ is positive. Hence, the drift in equation (B.6) is almost surely negative, and $\mathcal{E}_t$ is a super-martingale.

Using the super-martingale property, we find that $\mathbb{E}\left[\mathcal{E}_t\right] \leq \mathbb{E}\left[\mathcal{E}_0\right] = \mathbb{E}\left[D_h(x^\star, X_0 + e^{-\alpha_0}\nu_0) + e^{\beta_0}\left(f(X_0) - f(x^\star)\right)\right] = C_0$, where $C_0 \geq 0$. Using the definition of $\mathcal{E}$, and using the fact that $D_h \geq 0$ if h is convex, we obtain

$$e^{\beta_t}\mathbb{E}\left[(f(X_t) - f(x^\star))\right] \leq \mathbb{E}\left[D_h(x^\star, X_t + e^{-\alpha_t}\nu_t) + e^{\beta_t}\left(f(X_t) - f(x^\star)\right)\right] \leq C_0 + \mathbb{E}\left[\left[\nabla h(Y), Y\right]_t\right] . \tag{B.8}$$

Upper bound on the Quadratic Co-variation. Now we upper bound the quadratic co-variation term appearing on the right hand side of (B.8). Using the further change of variable $Z_t = \nabla h(Y_t)$, and noting that by the assumed convexity of h that $\nabla h^*(x) = (\nabla h)^{-1}(x)$, we get $[\nabla h(Y), Y]_t = [Z, \nabla h^*(Z)]_t$.

Assuming that ∇h is μ-strongly convex, we get that ∇h^* must have μ^{-1}-Lipschitz smooth gradients. This implies that (i) the eigenvalues of $\nabla^2 h^*$ must be bounded above by μ^{-1} (ii) from the Cauchy-Schwarz inequality, we have $\langle \nabla h^*(x) - \nabla h^*(y), x - y\rangle \leq \mu^{-1}\|x - y\|^2$. Using these two observations and writing out the expression for $[Z, \nabla h^*(Z)]_t$, we get

$$[Z, \nabla h^*(Z)]_t = \sum_{i,j=1}^{d} \frac{\partial^2 h(Y_t)}{\partial x_i \partial x_i} d\left[Y_i, Y_j\right]_t^c + \langle \Delta(\nabla h^*(Z)), \Delta Z_t\rangle \tag{B.9}$$

$$\leq \mu^{-1}[Z]_t . \tag{B.10}$$

Moreover, note that since $Z_t = \nabla h(X_t^{\nu^*} + e^{-\alpha_t}\nu_t^*)$ and since $\nabla h(X_t^{\nu^*})$ is a process of finite variation, the optimality dynamics (4.1) imply that $[Z]_t = [e^{-\gamma_t}\mathcal{M}]_t] = e^{-\gamma_t}[\mathcal{M}]_t$

Inserting the quadratic co-variation bound into equation (B.8) and using the super-martingale property, we obtain the final result

$$\begin{aligned}
\mathbb{E}\left[(f(X_t) - f(x^\star))\right] &\leq e^{-\beta_t}\left(C_0 + \frac{1}{2}\mathbb{E}\left[\left[\nabla h(X + e^{-\alpha_t}\nu), \nu\right]_t\right]\right) \\
&\leq e^{-\beta_t}\left(C_0 + \frac{1}{2}e^{-2\gamma_t}\mathbb{E}\left[\left[\mathcal{M}\right]_t\right]\right) \\
&\leq (C_0 + \frac{1}{2})e^{-\beta_t}\max\left\{1\,,\, e^{-2\gamma_t}\mathbb{E}\left[\left[\mathcal{M}\right]_t\right]\right\} \\
&= O\left(e^{-\beta_t}\max\left\{1\,,\, e^{-\beta_t + 2\gamma_t}\mathbb{E}\left[\left[\mathcal{M}\right]_t\right]\right\}\right) ,
\end{aligned}$$

as desired. □

Appendix C. Hamiltonian Representation of the Optimizer Dynamics

Just as in Hamiltonian classical mechanics, it is possible to express the optimality FBSDE of Theorem (4.1) with Hamiltonian equations of motion. We define the Hamiltonian $\mathcal{H}$ as the Legendre dual of $\mathcal{L}$ at, which can be written as

$$\mathcal{H}(t, X, p) = \langle p\,, \nu^*\rangle - \mathcal{L}(t, X, \nu^*) , \tag{C.1}$$

where $p = \frac{\partial \mathcal{L}}{\partial \dot{X}}$. Using the identity $D_h(x, y) = D_{h^*}(\nabla h(x), \nabla h(y))$, where h^* is the Legendre dual of h, and inverting the expression for $\frac{\partial \mathcal{L}}{\partial \dot{X}}$ in terms p, we may compute equation (C.1) as[2]

$$\mathcal{H}(t, X, p) = e^{\alpha_t + \gamma_t} D_{h^*}\left(\nabla h(X) + e^{-\gamma_t}p\,,\, \nabla h(X)\right) + e^{\gamma_t + \beta_t} f(X_t) . \tag{C.2}$$

Using this definition of $\mathcal{H}$, and using the FBSDE (4.1), we obtain the following equivalent representation for the dynamics of the optimizer.

Using the simple substitution $p_t = \left(\frac{\partial \mathcal{L}}{\partial \dot{X}}\right)_t$ and noting from equations (4.2) and (4.3) that

$$p_t = e^{\gamma_t}\left(\nabla h(X_t + e^{-\alpha_t}\nu_t^*) - \nabla h(X_t)\right) , \tag{C.3}$$

a straightforward computation applied to the definition of $\mathcal{H}$ shows that the dynamics of the FBSDE (4.1) admit the alternate Hamiltonian representation

$$dX_t = \left(\frac{\partial \mathcal{H}}{\partial p}\right)_t dt \;\;, \qquad dp_t = -\mathbb{E}\left[\left(\frac{\partial \mathcal{H}}{\partial X}\right)_t \Big| \mathcal{F}_t\right] dt - d\mathcal{M}_t \tag{C.4}$$

[2]See [28][Appendix B.4] for the full details of the computation.

along with the boundary condition $p_T = 0$. Noting here that (!) Finish this part.

Appendix D. Proof of Proposition 4.4

Proof. Computing the dynamics of $e^{\int_0^t \alpha_u \, du} \tilde{p}_t^{(0)}$, we find that

$$d(\tilde{p}_t^{(0)} e^{\int_0^t \alpha_u \, du}) = -e^{\int_0^t \alpha_u \, du + \alpha_t + \beta_t} \, \mathbb{E}\left[\nabla f\left(X_t\right) | \mathcal{F}_t\right] \, dt + e^{\int_0^t \alpha_u \, du} d\tilde{\mathcal{M}}_t \, , \tag{D.1}$$

along with the same boundary condition. Writing out the solution explicitly, and multiplying both sides by $e^{-\int_0^t \alpha_u \, du}$ we get that

$$\tilde{p}_t^{(0)} = \mathbb{E}\left[\int_t^T e^{\int_t^u \alpha_s \, ds + \alpha_u + \beta_u} \, \nabla f\left(X_u\right) \, du - e^{\delta_T - \gamma_T + \int_t^T \alpha_s \, ds} \, \nabla f(X_T^{\nu^*}) \Big| \mathcal{F}_u \right] \, . \tag{D.2}$$

Noting that under the assumption of the ideal scaling conditions, we have $\gamma_t - \gamma_0 = \int_0^t e^{\alpha_t}$, we get that $e^{\int_t^T \alpha_s \, ds - \gamma_T} = e^{-\gamma_t}$, yielding the desired expression.

By the assumption that δ, γ are continuous over $[0, T]$ and that $\mathbb{E}\|f(x)\|^2\| < \infty$, the right part of (D.2) is bounded. Now note that the integral on the left side of (D.2) is upper bounded for all T by the integral provided in the integrability condition of Proposition 4.4, and therefore this condition is a sufficient condition for the expression (D.2) to be finite and well-defined.

□

References

[1] Laurence Aitchison, *A unified theory of adaptive stochastic gradient descent as bayesian filtering*, arXiv preprint arXiv:1807.07540 (2018).

[2] Amir Beck and Marc Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters **31** (2003), no. 3, 167–175.

[3] Alain Bensoussan, *Stochastic control of partially observable systems*, Cambridge University Press, 2004.

[4] René Carmona, *Lectures on bsdes, stochastic control, and stochastic differential games with financial applications*, Vol. 1, SIAM, 2016.

[5] Philippe Casgrain and Sebastian Jaimungal, *Algorithmic trading with partial information: A mean field game approach*, arXiv preprint arXiv:1803.04094 (2018).

[6] ———, *Mean-field games with differing beliefs for algorithmic trading*, arXiv preprint arXiv:1810.06101 (2018).

[7] ———, *Trading algorithms with learning in latent alpha models*, arXiv preprint arXiv:1806.04472 (2018).

[8] Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile, *On the generalization ability of on-line learning algorithms*, IEEE Transactions on Information Theory **50** (2004), no. 9, 2050–2057.

[9] John Duchi, Elad Hazan, and Yoram Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, Journal of Machine Learning Research **12** (2011), no. Jul, 2121–2159.

[10] Vineet Gupta, Tomer Koren, and Yoram Singer, *A unified approach to adaptive regularization in online and stochastic optimization*, arXiv preprint arXiv:1706.06569 (2017).

[11] Jouni Hartikainen and Simo Särkkä, *Kalman filtering and smoothing solutions to temporal gaussian process regression models*, Machine learning for signal processing (mlsp), 2010 ieee international workshop on, 2010, pp. 379–384.

[12] Jean Jacod and Albert Shiryaev, *Limit theorems for stochastic processes*, Vol. 288, Springer Science & Business Media, 2013.

[13] Svetlana Janković, Miljana Jovanović, and Jasmina Djordjević, *Perturbed backward stochastic differential equations*, Mathematical and Computer Modelling **55** (2012), no. 5-6, 1734–1745.

[14] Toni Karvonen and Simo Sarkkä, *Approximate state-space gaussian processes via spectral transformation*, Machine learning for signal processing (mlsp), 2016 ieee 26th international workshop on, 2016, pp. 1–6.

[15] Diederik P Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980 (2014).

[16] James Lucas, Shengyang Sun, Richard Zemel, and Roger Grosse, *Aggregated momentum: Stability through passive damping*, arXiv preprint arXiv:1804.00325 (2018).

[17] Jin Ma, J-M Morel, and Jiongmin Yong, *Forward-backward stochastic differential equations and their applications*, Springer Science & Business Media, 1999.

[18] Arkadii Semenovich Nemirovsky and David Borisovich Yudin, *Problem complexity and method efficiency in optimization.* (1983).

[19] Yu Nesterov, *A method of solving a convex programming problem with convergence rate o (1/k2)*, Sov. math. dokl.

[20] Etienne Pardoux and Shanjian Tang, *Forward-backward stochastic differential equations and quasilinear parabolic pdes*, Probability Theory and Related Fields **114** (1999), no. 2, 123–150.

[21] Boris T Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Computational Mathematics and Mathematical Physics **4** (1964), no. 5, 1–17.

[22] Herbert Robbins and Sutton Monro, *A stochastic approximation method*, The annals of mathematical statistics (1951), 400–407.

[23] Sebastian Ruder, *An overview of gradient descent optimization algorithms*, arXiv preprint arXiv:1609.04747 (2016).

[24] Simo Särkkä and Jouni Hartikainen, *Infinite-dimensional kalman filtering approach to spatio-temporal gaussian process regression*, International conference on artificial intelligence and statistics, 2012, pp. 993–1001.

[25] Simo Särkkä and Robert Piché, *On convergence and accuracy of state-space approximations of squared exponential covariance functions*, Machine learning for signal processing (mlsp), 2014 ieee international workshop on, 2014, pp. 1–6.

[26] Ramon Van Handel, *Stochastic calculus, filtering, and stochastic control*, Course notes., URL http://www. princeton. edu/˜ rvan/acm217/ACM217. pdf (2007).

[27] James Vuckovic, *Kalman gradient descent: Adaptive variance reduction in stochastic optimization*, arXiv preprint arXiv:1810.12273 (2018).

[28] Andre Wibisono, Ashia C Wilson, and Michael I Jordan, *A variational perspective on accelerated methods in optimization*, Proceedings of the National Academy of Sciences **113** (2016), no. 47, E7351–E7358.