SIAM J. CONTROL OPTIM.
Vol. 60, No. 6, pp. 3191-3216

EXPLORATORY HJB EQUATIONS AND THEIR CONVERGENCE*

WENPIN TANG†, YUMING PAUL ZHANG‡, AND XUN YU ZHOU†

Abstract. We study the exploratory Hamilton–Jacobi–Bellman (HJB) equation arising from the entropy-regularized exploratory control problem, which was formulated by Wang, Zariphopoulou, and Zhou (J. Mach. Learn. Res., 21 (2020), 198) in the context of reinforcement learning in continuous time and space. We establish the well-posedness and regularity of the viscosity solution to the equation, as well as the convergence of the exploratory control problem to the classical stochastic control problem when the level of exploration decays to zero. We then apply the general results obtained to the exploratory temperature control problem, which was introduced by Gao, Xu, and Zhou (SIAM J. Control Optim., 60 (2022), pp. 1250–1268) to design an endogenous temperature schedule for simulated annealing in the context of nonconvex optimization. We derive an explicit rate of convergence for this problem as exploration diminishes to zero, and find that the stationary distribution of the optimally controlled process exists, which is however neither a Dirac mass on the global optimum nor a Gibbs measure.

Key words. HJB equations, stochastic control, partial differential equations, reinforcement learning, exploratory control, entropy regularization, simulated annealing, overdamped Langevin equation

MSC codes. 35F21, 60J60, 93E15, 93E20

DOI. 10.1137/21M1448185

1. Introduction.

Reinforcement learning (RL) is an active subarea of machine learning. RL research has predominantly focused on Markov decision processes in discrete time and space; see [29] for a systematic account of the theory and applications, as well as a detailed description of bibliographical and historical development of the field. Wang, Zariphopoulou, and Zhou [32] are probably the first to formulate and develop an entropy-regularized, exploratory control framework for RL in continuous time with continuous feature (state) and action (control) spaces. In this framework, stochastic relaxed control, a measure-valued process, is employed to represent exploration through randomization, capturing the notion of "trial and error" which is the core of RL. Entropy of the control is incorporated explicitly as a regularization term in the objective function to encourage exploration, with a weight parameter $\lambda > 0$ on the entropy to gauge the tradeoff between exploitation (optimization) and exploration (randomization). This exploratory formulation has been extended to other settings and used to solve applied problems; see, e.g., [13] and [17] to mean-field games, and [33] to Markowitz mean-variance portfolio optimization. [15] apply the same formulation to temperature control of Langevin diffusions arising from simulated annealing

* Received by the editors September 23, 2021; accepted for publication (in revised form) August 1, 2022; published electronically November 11, 2022.
https://doi.org/10.1137/21M1448185
Funding: The first author gratefully acknowledges financial support through NSF grants DMS-2113779 and DMS-2206038, and through a startup grant at Columbia University. The third author gratefully acknowledges financial support through a startup grant at Columbia University and through the Nie Center for Intelligent Asset Management.

† Department of Industrial Engineering and Operations Research, Columbia University, New York, NY USA 10027 (wt2319@columbia.edu, xz2574@columbia.edu).

‡ Department of Mathematics, University of California, San Diego, CA USA (yzhangpaul@ucsd.edu).

for nonconvex optimization. The problem itself is not directly related to RL; however, the authors take the same idea of "exploration through randomization" and invoke exploratory controls to smooth out the highly unstable yet theoretically optimal bang-bang control. For more literature review on the exploratory control, see [37].

Wang, Zariphopoulou, and Zhou [32] derive the following Hamilton–Jacobi–Bellman (HJB) partial differential equation (PDE) associated with the exploratory control problem, parameterized by the weight parameter $\lambda > 0$:

$$-\rho v_\lambda(x) + \lambda \ln \int_{\mathcal{U}} \exp\left(\frac{1}{\lambda}\Big[h(x,u) + b(x,u)\cdot \nabla v_\lambda(x) + \frac{1}{2}\mathrm{Tr}(\sigma(x,u)\sigma(x,u)^T \nabla^2 v_\lambda(x))\Big]\right) du = 0. \tag{1.1}$$

This equation, called the *exploratory HJB equation*, appears to be characteristically different from the HJB equation corresponding to a classical stochastic control problem. Among other things, (1.1) does not involve the supremum operator in the control variable typically appearing in a classical HJB equation. This is because the supremum is replaced by a distribution among controls in the exploratory formulation. Wang Zariphopoulou, and Zhou [32] do not study this general equation in terms of its well-posedness (existence and uniqueness of the viscosity solution), regularity, stability in λ, or the convergence when $\lambda \to 0^+$. They do, however, solve the important linear-quadratic (LQ) case where the exploratory HJB equation can be solved explicitly, leading to the optimal distribution for exploration being a Gaussian distribution. Wang and Zhou [33] apply this result to a continuous-time Markowitz portfolio selection problem which is inherently LQ.

The goal of this paper is to study the general exploratory HJB equations beyond the LQ setting. We first analyze a class of elliptic PDEs under fairly general assumptions on the coefficients (Theorems 3.6 and 3.7). The application of the general results obtained to the exploratory HJB equations allows us to identify the assumptions needed, to derive the well-posedness of viscosity solutions and their regularity, and to establish a connection between the exploratory control problem and the classical stochastic control problem (Theorems 3.9 and 3.10). More specifically to the last point, we show that as the exploration weight decays to zero, the value function of the former converges to that of the latter. This result, which extends [32] to the general setting, is important for RL especially in terms of finding the regret bound (or the cost of exploration as termed in [32]). As a passing note, our analysis for the general class of fully nonlinear elliptic PDEs may be of independent interest to the PDE community.

In the second part of this paper, we focus on a special exploratory HJB equation resulting from the exploratory temperature control problem of the Langevin diffusions. The latter problem was introduced by Gao, Xu, and Zhou [15] aiming at designing a state-dependent temperature schedule for simulated annealing (SA). To provide a brief background (see [15] for more details), one of the central problems in continuous optimization is to escape from saddle points and local minima, and to find a global minimum of a nonconvex function $f : \mathbb{R}^d \to \mathbb{R}$. Applying the SA technique to the gradient descent algorithm consists of adding a sequence of independent Gaussian noises, scaled by "temperature" parameters controlling the level of noises. The continuous version of the SA algorithm is governed by the following stochastic differential equation (SDE),

$$dX_t = -\nabla f(X_t)dt + \sqrt{2\beta_t}dB_t, \quad X_0 = x, \tag{1.2}$$

where $(B_t,\, t \geq 0)$ is a d-dimensional Brownian motion, and the temperature schedule $(\beta_t,\, t \geq 0)$ is a stochastic process. If $\beta_t \equiv \beta$ is constant in time, then (1.2) is the well-known overdamped Langevin equation whose stationary distribution is the Gibbs measure $\mathcal{G}_\beta(dx) \propto \exp(-f(x)/\beta)dx$ (f is called the landscape and β the temperature).

When allowing $(\beta_t,\, t \geq 0)$ to be a stochastic process, we have naturally a stochastic control problem in which one controls the dynamics (1.2) through this temperature process in order to achieve the highest efficiency in optimizing f. Gao,Xu, and Zhou [15] find that the optimal control of this problem is of bang-bang type: the temperature process switches between two extremum points in the search interval. Such a bang-bang solution is almost unusable in practice since it is highly sensitive to errors. Moreover, in the present paper we discover that the optimal state process under the bang-bang control may even not be well-posed in dimensions $d \geq 3$ (section 4.1). These observations support the entropy-regularized exploratory formulation of temperature control proposed by [15], not so much from a learning perspective, but from a desire of smoothing out the bang-bang control.

The results for the general exploratory HJB equations apply readily to the temperature control setting in terms of the well-posedness, regularity, and convergence (Corollaries 4.3 and 4.7). Moreover, due to the special structure of the controlled dynamics (1.2), we are able to derive an explicit convergence rate of $\lambda \ln(1/\lambda)$ for the exploratory temperature control problem as λ tends to zero (Theorem 4.6). Finally, we consider the long time behavior of the associated optimally controlled process and show that it will not converge to the global minimum of f nor any Gibbs measure with landscape f (Theorem 4.9). The first property is indeed preferred from an exploration point of view because exploration is meant to involve as many states as possible instead of focusing only on the single state of the minimizer. The second property hints at the possibility of more variety of target measures other than Gibbs measures for SA. Finally, while the main focus of the paper is on problems in the infinite time horizon, our results also carry over to the exploratory control problem in a finite time horizon (Theorem 5.2).

The remainder of the paper is organized as follows. In section 2, we provide some background on the exploratory control framework and present the corresponding exploratory HJB equation. In section 3, we investigate the exploratory HJB equation and establish general results in terms of its well-posedness, regularity, and convergence. We also identify the value function of the exploratory control problem as the unique solution to the exploratory HJB equation. In section 4, we apply the general results to the exploratory temperature control problem, derive an explicit convergence rate, and study the long time behavior of the associated optimal state process. In section 5 we consider the exploratory control problem in a finite time horizon. Finally, section 6 concludes.

2. Background and problem formulation. In this section, we provide some background on the exploratory control problem that is put forth in [32].

Below we collect some notations that will be used throughout this paper.

- For $x, y \in \mathbb{R}^d$, $x{\cdot}y$ denotes the inner product between x and y, $|x| = \sqrt{\sum_{i=1}^d x_i^2}$ denotes the Euclidean norm of x, $B_R = \{x : |x| \leq R\}$ denotes the Euclidean ball of radius R centered at 0, and $|x|_{\max} = \max_{1\leq i\leq d} |x_i|$ denotes the max norm of x.
- For a square matrix $X = (X_{ij}) \in \mathbb{R}^{d\times d}$, X^T denotes its transpose, $\mathrm{Tr}(X)$ its trace, $|X|$ its spectral norm, and $|X|_{\max} = \max_{1\leq i,j\leq d} |X_{ij}|$ its max norm.

Moreover, $\mathcal{S}^d = \{X \in \mathbb{R}^{d\times d} : X^T = X\}$ denotes the set of $d \times d$ symmetric matrices with the spectral norm.

- Let $\mathcal{O} \subseteq \mathbb{R}^d$ be open. For a function $f : \mathcal{O} \to \mathbb{R}$, ∇f, $\nabla^2 f$, and $\Delta f = \mathrm{Tr}(\nabla^2 f)$ denote, respectively, its gradient, Hessian, and Laplacian.
- For a bounded function $f : \mathcal{O} \to \mathbb{R}$, $||f||_{L^\infty(\mathcal{O})} = \sup_{x\in\mathcal{O}} |f(x)|$ denotes the sup norm of f.
- A function $f \in \mathcal{C}^k(\mathcal{O})$, or simply $f \in \mathcal{C}^k$ if it is k-time continuously differentiable. The $\mathcal{C}^k$ norm is given by $||f||_{\mathcal{C}^k} = \max_{|\beta|\le k} \sup_{x\in\mathcal{O}} |\nabla^\beta f(x)|$, where $\nabla^\beta f(x) = \frac{\partial^{|\beta|} f}{\partial x_1^{\beta_1}\cdots\partial x_d^{\beta_d}}(x)$ with $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{N}^d$ and $|\beta| = \sum_{i=1}^d \beta_i$.
- A function $f \in \mathcal{C}^{k,\alpha}(\mathcal{O})$, or simply $f \in \mathcal{C}^{k,\alpha}$ $(0 < \alpha < 1)$, if it is k-time continuously differentiable and its kth derivatives of f are α-Hölder continuous. The $\mathcal{C}^{k,\alpha}$ norm is given by $||f||_{\mathcal{C}^{k,\alpha}} = \max_{|\beta|\le k} \sup_{x\in\mathcal{O}} |\nabla^\beta f(x)| + \max_{|\beta|=k} \sup_{x\neq y\in\mathcal{O}} \frac{|\nabla^\beta f(x) - \nabla^\beta f(y)|}{|x-y|^\alpha}$.
- For two probability measures $\mathbb{P}$ and $\mathbb{Q}$, $||\mathbb{P} - \mathbb{Q}||_{TV} = \sup_A |\mathbb{P}(A) - \mathbb{Q}(A)|$ denotes the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$.

2.1. Classical control problem. Let $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\}_{t\ge 0})$ be a filtered probability space on which we define a d-dimensional $\mathcal{F}_t$-adapted Brownian motion $(B_t,\, t \ge 0)$. Let $\mathcal{U}$ be a generic action/control space, and $u = (u_t,\, t \ge 0)$ be a control which is an $\mathcal{F}_t$-adapted process taking values in $\mathcal{U}$.

The classical stochastic control problem is to control the state variable $X_t \in \mathbb{R}^d$, whose dynamics is governed by the SDE

$$dX_t^u = b(X_t^u, u_t)dt + \sigma(X_t^u, u_t)dB_t, \quad X_0^u = x, \tag{2.1}$$

where $b : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d$ is the drift, and $\sigma : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^{d\times d}$ is the covariance matrix of the state variable. Here the superscript 'u' in X_t^u emphasizes the dependence of the state variable on the control u. The goal of the control problem is to maximize the total discounted reward, leading to the (optimal) value function

$$v(x) = \sup_{u\in\mathcal{A}_0(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t} h(X_t^u, u_t)dt \,\middle|\, X_0^u = x\right], \tag{2.2}$$

where $h : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}$ is a reward function, $\rho > 0$ is the discount factor, and $\mathcal{A}_0(x)$ denotes the set of admissible controls which may depend on the initial state value $X_0^u = x$.

By a standard dynamic programming argument, the HJB equation associated with problem (2.2) is

$$-\rho v(x) + \sup_{u\in\mathcal{U}} \left[h(x,u) + b(x,u)\cdot\nabla v(x) + \frac{1}{2}\mathrm{Tr}(\sigma(x,u)\sigma(x,u)^T\nabla^2 v(x))\right] = 0. \tag{2.3}$$

In the classical stochastic control setting, the functional forms of h, b, σ are given and known. It is known that a suitably smooth solution to the HJB equation (2.3) gives the value function (2.2). Further, the optimal control is represented as a deterministic mapping from the current state to the action/control space: $u_t^* = u^*(X_t^*)$. The mapping u^* is called an optimal feedback control, which is derived offline from the "$\sup_{u\in\mathcal{U}}$" term in (2.3). This procedure of obtaining the optimal feedback control is called the verification theorem. The corresponding optimally controlled process $(X_t^*,\, t \ge 0)$ is governed by the SDE,

$$dX_t^* = b(X_t^*, u^*(X_t^*))dt + \sigma(X_t^*, u^*(X_t^*))dB_t, \quad X_0^* = x,$$

provided that it is well-posed (i.e., it has a unique weak solution). See, e.g., [14, 36] for detailed accounts of the classical stochastic control theory.

2.2. Exploratory control problem. In the RL setting, the model parameters are unknown, i.e., the functions h, b, σ are not known. Thus, one needs to explore and learn the optimal controls through repeated trials and errors. Inspired by this, [32] model exploration by a probability distribution of controls $\pi = (\pi_t(\cdot),\, t \geq 0)$ over the control space $\mathcal{U}$ from which each trial is sampled. The exploratory state dynamics is

$$dX_t^\pi = \widetilde{b}(X_t^\pi, \pi_t)dt + \widetilde{\sigma}(X_t^\pi, \pi_t)dB_t, \quad X_0^\pi = x, \tag{2.4}$$

where the coefficients $\widetilde{b}(\cdot,\cdot)$ and $\widetilde{\sigma}(\cdot,\cdot)$ are defined by

$$\widetilde{b}(x,\pi) := \int_{\mathcal{U}} b(x,u)\pi(u)du, \quad \widetilde{\sigma}(x,\pi) := \left(\int_{\mathcal{U}} \sigma(x,u)\sigma(x,u)^T\pi(u)du\right)^{\frac{1}{2}}$$

for $(x,\pi) \in \mathbb{R}^d \times \mathcal{P}(\mathcal{U})$ with $\mathcal{P}(\mathcal{U})$ being the set of absolutely continuous probability density functions on $\mathcal{U}$. The distributional control $\pi = (\pi_t(\cdot),\, t \geq 0)$ is also known as the relaxed control, and a classical control $u = (u_t,\, t \geq 0)$ is a special relaxed control when $\pi_t(\cdot)$ is taken as the Dirac mass at u_t.

The exploratory control problem is an optimization problem similar to (2.2) but under relaxed controls. Moreover, to encourage exploration, Shannon's entropy is added to the objective function as a regularization term:

$$v_\lambda(x) = \sup_{\pi\in\mathcal{A}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\int_{\mathcal{U}} h(X_t^\pi, u)\pi_t(u)du - \lambda\int_{\mathcal{U}} \pi_t(u)\ln\pi_t(u)du\right)dt \,\middle|\, X_0^u = x\right], \tag{2.5}$$

where $\lambda > 0$ is a weight parameter controlling the level of exploration (also called the temperature parameter), and $\mathcal{A}(x)$ is the set of admissible distributional controls specified by the following definition.

Definition 2.1. *We say a density-function-valued stochastic process $\pi = (\pi_t(\cdot),\, t \geq 0)$, defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\}_{t\geq 0})$ along with a d-dimensional $\mathcal{F}_t$-adapted Brownian motion $(B_t,\, t \geq 0)$, is an admissible distributional (or exploratory) control, denoted by $\pi \in \mathcal{A}(x)$, if*

(i) *for each $t \geq 0$, $\pi_t(\cdot) \in \mathcal{P}(\mathcal{U})$ a.s.;*
(ii) *for any Borel subset $A \subset \mathcal{U}$, the process $(t,\omega) \to \int_A \pi_t(u,\omega)du$ is $\mathcal{F}_t$-progressively measurable;*
(iii) *the SDE (2.4) has solutions on the same filtered probability space whose distributions are all identical.*

Now we quickly review a formal derivation of the solution to the exploratory control problem (2.4)–(2.5), following [32]. By dynamic programming, the HJB equation to (2.4)–(2.5) is

$$\begin{aligned} -\rho v_\lambda(x) + \sup_{\pi\in\mathcal{P}(\mathcal{U})} \int_{\mathcal{U}} \Big(& h(x,u) + b(x,u)\cdot\nabla v_\lambda(x) \\ &+ \frac{1}{2}\mathrm{Tr}(\sigma(x,u)\sigma(x,u)^T\nabla^2 v_\lambda(x)) - \lambda\ln\pi(u)\Big)\pi(u)du = 0. \end{aligned} \tag{2.6}$$

Then, through the same verification theorem argument, the optimal feedback control is

(2.7)
$$\pi^*(u,x)=\frac{\exp\left(\frac{1}{\lambda}\left[h(x,u)+b(x,u)\cdot\nabla v_\lambda(x)+\frac{1}{2}\mathrm{Tr}(\sigma(x,u)\sigma(x,u)^T\nabla^2 v_\lambda(x))\right]\right)}{\int_{\mathcal{U}}\exp\left(\frac{1}{\lambda}\left[h(x,u)+b(x,u)\cdot\nabla v_\lambda(x)+\frac{1}{2}\mathrm{Tr}(\sigma(x,u)\sigma(x,u)^T\nabla^2 v_\lambda(x))\right]\right)du},$$
which is the Boltzmann distribution or a Gibbs measure. By injecting (2.7) into (2.6), we get the nonlinear elliptic PDE (1.1), or the exploratory HJB equation. Note that this equation is parameterized by the weight parameter $\lambda>0$.

Applying the feedback control (2.7) to the state dynamics (2.4), we obtain the optimally controlled dynamics

$$dX_t^{\lambda,*}=\widetilde{b}(X_t^{\lambda,*},\pi^*(\cdot,X_t^{\lambda,*}))dt+\widetilde{\sigma}(X_t^{\lambda,*},\pi^*(\cdot,X_t^{\lambda,*}))dB_t, \tag{2.8}$$
provided that it is well-posed, i.e., it has a weak solution which is unique in distribution. This condition is satisfied if $b(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ are measurable and bounded, $x\to\sigma(x,\cdot)$ is continuous, and $\sigma(\cdot,\cdot)$ is strictly elliptic in the sense that $\sigma(\cdot,\cdot)\sigma(\cdot,\cdot)^T\geq\Lambda I$; see, e.g., [28] for discussions on the well-posedness of SDEs. The optimal distributional control is then $\pi_t^{\lambda,*}(\cdot)=\pi^*(\cdot,X_t^{\lambda,*})$, $t\geq 0$.

The exploratory HJB equation (1.1) is a new type of PDE in control theory, which begs a number of questions. [25] considered the exploratory control problem in bounded domains, while our interest is in the whole space $\mathbb{R}^d$ motivated by applications. The first question is, naturally, its well-posedness (existence and uniqueness) in a certain sense. The second question is its dependence and convergence in $\lambda>0$. In practice, this parameter is often set to be small. Thus, we are interested in the limit of the solution to (1.1) as $\lambda\to 0^+$, along with its convergence rate. We will answer these questions in the following two sections.

3. Analysis of exploratory HJB equation. In this section, we study the exploratory HJB equation (1.1) under some general assumptions on the functions $h(\cdot,\cdot),b(\cdot,\cdot),\sigma(\cdot,\cdot)$. For a concise analysis it is advantageous to analyze the general fully nonlinear elliptic PDEs of the form
$$F(\nabla^2 v,\nabla v,v,x)=0\quad\text{ in }\mathbb{R}^d, \tag{3.1}$$
and then apply the results obtained to (1.1).

3.1. General results on second order elliptic equations. The standard references for second order elliptic PDEs are [16, 7]. Here we recall some definitions and useful results.

Consider the general fully nonlinear equations (3.1). We make the following assumptions on the operator $F:\mathcal{S}^d\times\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}^d\to\mathbb{R}$:

(a) F is continuous in all its variables, and for each $r\geq 1$ there exist $\gamma_r,\underline{\gamma_r}>0$ such that for any $x,y\in B_r$ and $(X,p,q,s)\in\mathcal{S}^d\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}$,
$$|F(X,p,s,x)-F(X,q,s,y)|\leq\gamma_r|x-y|(1+|p|+|q|+|X|)+\underline{\gamma_r}|p-q|,$$
$$|F(0,0,0,x)|\leq\gamma_r.$$

(b) There exist $\Lambda_2>\Lambda_1>0$ such that for any $P\in\mathcal{S}^d$ positive semidefinite, and any $(X,p,s,x)\in\mathcal{S}^d\times\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}^d$,
$$\Lambda_2\mathrm{Tr}(P)\geq F(X,p,s,x)-F(X+P,p,s,x)\geq\Lambda_1\mathrm{Tr}(P).$$

(c) There exists $\rho>0$ such that for all $(X,p,x)\in\mathcal{S}^d\times\mathbb{R}^d\times\mathbb{R}^d$ and $t\geq s$,
$$F(X,p,t,x)-F(X,p,s,x)\geq\rho(t-s).$$

These assumptions are standard (see [18, 9]), and guarantee the existence and uniqueness of the viscosity solution to (3.1) in a bounded domain with a Dirichlet boundary condition. The proof is given by Perron's method and the comparison principle. Note that there exist weaker conditions than the ones stated above to ensure the well-posedness of (3.1) in bounded domains; however, assumptions (a)–(c) are simpler and sufficient for our purpose.

Now we recall the definition of viscosity solutions to (3.1).

DEFINITION 3.1. *Let Ω be an open set in $\mathbb{R}^d$.*

(i) *We say an upper semicontinuous (resp., lower semicontinuous) function $v : \Omega \to \mathbb{R}$ is a subsolution (resp., supersolution) to* (3.1) *if the following holds: for any smooth function ϕ in Ω such that $v - \phi$ has a local maximum (resp., minimum) at $x_0 \in \Omega$, we have*

$$F(\nabla^2\phi, \nabla\phi, v(x_0), x_0) \le 0 \quad (\text{resp., } F(\nabla^2\phi, \nabla\phi, v(x_0), x_0) \ge 0).$$

(ii) *We say a continuous function $v : \Omega \to \mathbb{R}$ is a (viscosity) solution to* (3.1) *if it is both a subsolution and a supersolution.*

Throughout this paper, by a solution of a PDE we mean a *viscosity* solution unless otherwise stated.

Assume that there are a set of functions defined on Ω: $\{v_\varepsilon(x), \varepsilon > 0\}$. Recall the definition of half-relaxed limits: $v^*(x) := \limsup_{\substack{\Omega \ni x' \to x, \\ \varepsilon \to 0}} v_\varepsilon(x')$ and $v_*(x) := \liminf_{\substack{\Omega \ni x' \to x, \\ \varepsilon \to 0}} v_\varepsilon(x')$. Clearly, v^* is upper semicontinuous and v_* is lower semicontinuous. It is known that subsolutions and supersolutions are stable under the half-relaxed limit operations; see [9].

LEMMA 3.2. *Let $\Omega \subseteq \mathbb{R}^d$ be open, $\{F_\lambda, \lambda > 0\}$ be a set of operators satisfying the assumptions* (a)–(c) *with the same constants. Suppose that F_λ converges locally uniformly in all its variables to an operator $\bar{F}$ as $\lambda \to 0^+$. Then*

(i) *if v_λ is a sequence of bounded subsolutions to $F_\lambda(\nabla^2 v_\lambda, \nabla v_\lambda, v_\lambda, \cdot) \le 0$ in Ω for some $\lambda \to 0^+$,, then their upper half-relaxed limit v^* is a subsolution to*

$$\bar{F}(\nabla^2 v^*, \nabla v^*, v^*, \cdot) \le 0 \quad \text{in } \Omega;$$

(ii) *if v_λ is a sequence of bounded supersolutions to $F_\lambda(\nabla^2 v_\lambda, \nabla v_\lambda, v_\lambda, \cdot) \ge 0$ in Ω for some $\lambda \to 0^+$, then their lower half-relaxed limit v_* is a supersolution to*

$$\bar{F}(\nabla^2 v_*, \nabla v_*, v_*, \cdot) \ge 0 \quad \text{in } \Omega.$$

Next we consider the regularity of solutions to (3.1). We need the following additional assumption on the operator F.

DEFINITION 3.3. *We say that an operator $F = F(X, p, s, x)$ is concave in X if for any $M, N \in \mathcal{S}^d, p, x \in \mathbb{R}^d$, and $s \in \mathbb{R}$ we have*

$$-\frac{\partial^2 F(M, p, s, x)}{\partial M_{ij} \partial M_{kl}} N_{ij} N_{kl} \le 0,$$

where the derivative and the inequality are in the sense of distribution.

The following result concerns higher regularity of bounded solutions to concave operators; see, e.g., [7] and [22]. As a consequence, viscosity solutions to concave operators are classical solutions.

LEMMA 3.4 (Theorems 2.1 and 2.6 [22]). *Assume that $F = F(X, p, s, x)$ satisfies (a)–(c), and let $R_2 > R_1 > 0$. If v is a bounded viscosity solution to the equation $F(\nabla^2 v, \nabla v, v, x) = 0$ in B_{R_2}, then v is $\mathcal{C}^{1,\alpha}$ in B_{R_1}. Moreover if F is concave in X, then v is $\mathcal{C}^{2,\alpha}$ in B_{R_1}. The upper bounds for $||v||_{\mathcal{C}^{1,\alpha}(B_{R_1})}$ or $||v||_{\mathcal{C}^{2,\alpha}(B_{R_1})}$ depend only on the constants in assumptions (a)–(c), R_1, R_2, and $\|v\|_{L^\infty(B_{R_2})}$.*

Finally, we prove a comparison principle for solutions to (3.1), where the operator F is assumed to have a certain subquadratic growth in x in the whole domain $\mathbb{R}^d$. This comparison principle will be used to prove the uniqueness of the solution to the exploratory HJB equation (1.1) under some assumptions on $h(\cdot,\cdot), b(\cdot,\cdot), \sigma(\cdot,\cdot)$.

LEMMA 3.5 (Comparison principle in $\mathbb{R}^d$). *Assume that F satisfies (a)–(c) with $\underline{\gamma_r} > 0$ such that*

$$\limsup_{r\to\infty} \underline{\gamma_r}/r = 0. \tag{3.2}$$

Let v_1 and v_2 be locally uniformly bounded and be, respectively, a subsolution and a supersolution to (3.1) in $\mathbb{R}^d$ such that

$$\limsup_{|x|\to\infty} \frac{v_1(x) - v_2(x)}{|x|^2} \le 0. \tag{3.3}$$

Then $v_1 \le v_2$ in $\mathbb{R}^d$.

Proof. Our proof relies on a classical comparison principle of [18] for elliptic PDEs in a *bounded* domain.

It follows from (3.2) that there exists $C > 0$ such that for all $r \ge 0$,

$$(C + r^2)\rho \ge 2\underline{\gamma_r} r. \tag{3.4}$$

Set $C' := C + 2d\Lambda_2\rho^{-1}$, and for any small $\varepsilon > 0$, define $v^\varepsilon(x) := v_2(x) + \varepsilon(C' + |x|^2)$. We claim that v^ε is a supersolution to (3.1) in $\mathbb{R}^d$. Indeed, assume that there is $\varphi \in \mathcal{C}^\infty(\mathbb{R}^d)$ such that $v^\varepsilon - \varphi$ has a local minimum at $x_0 \in \mathbb{R}^d$. Then $v - \varphi^\varepsilon$ with $\varphi^\varepsilon := \varphi - \varepsilon(C' + |x|^2)$ has a local minimum at x_0. Using the facts that v_2 is a supersolution and F satisfies (a)–(c), we get by (3.4) that

$$\begin{aligned}
&F(\nabla^2\varphi, \nabla\varphi, v^\varepsilon(x_0), x_0)\\
&\ge F(\nabla^2\varphi^\varepsilon, \nabla\varphi^\varepsilon, v_2(x_0), x_0) - 2d\Lambda_2\varepsilon + \rho(C' + |x_0|^2)\varepsilon - 2\underline{\gamma_{|x_0|}}|x_0|\varepsilon\\
&\ge (C + |x_0|^2)\rho\varepsilon - 2\underline{\gamma_{|x_0|}}|x_0|\varepsilon \ge 0.
\end{aligned}$$

Hence v^ε is a supersolution.

Next, due to (3.3), there exists $R_\varepsilon > 0$ such that $v^\varepsilon(x) \ge v_1(x)$ for all $|x| \ge R_\varepsilon$. Therefore applying [18, Theorem III.1] to v_1, v^ε in the bounded domain B_{R_ε} yields

$$v^\varepsilon(x) \ge v_1(x) \quad \text{for all } x \in B_{R_\varepsilon}.$$

Taking $\varepsilon \to 0$ leads to $v_2 \ge v_1$ in $\mathbb{R}^d$. ∎

The above proof of Lemma 3.5 follows rather standard lines. The comparison principle (and the well-posedness) for unbounded solutions to nonlinear elliptic equations in unbounded domains do exist in the literature; see, e.g., [1, 8, 9, 19]. However, those results do not apply to the problem in which we are interested. In particular, none of these results covers the cases of unbounded $b(\cdot,\cdot)$ and/or F being inhomogeneous in X, inherent in the exploratory control problem.

3.2. Well-posedness and stability. In this subsection, we prove the well-posedness of solutions of subquadratic growth to (3.1). We need some assumptions on $\gamma_r, \underline{\gamma_r}$. Let $\gamma : (0,\infty) \to (0,\infty)$ be $\mathcal{C}^2$. Setting $\gamma_r := \gamma(r), \gamma_r{}' := \gamma'(r), \gamma_r{}'' := \gamma''(r)$, we assume that

$$\gamma_r{}' \geq 0 \quad \text{and} \quad \limsup_{r\to\infty} \frac{\gamma_r{}'}{r} + \frac{\gamma_r{}' + |\gamma_r{}''|}{\gamma_r} = 0. \tag{3.5}$$

Note that $\limsup_{r\to\infty} \frac{\gamma_r{}'}{r} = 0$ implies $\limsup_{r\to\infty} \frac{\gamma_r}{r^2} = 0$. So this γ_r represents a rate of subquadratic growth. For instance, we can take $\gamma_r = C(1+r^a)$ or $C(1+r^a \ln(1+r))$ with $a \in [0,2), C > 0$. The assumption on $\gamma_r{}''$ avoids large oscillations of γ_r when $r \to \infty$.

THEOREM 3.6. *The following hold:*

(a) *Assume that (a)–(c) hold with γ_r satisfying* (3.5) *and $\underline{\gamma_r}$ satisfying*

$$\limsup_{r\to\infty} (\underline{\gamma_r} - \gamma_r/r) < \infty. \tag{3.6}$$

Then there exists a unique solution v of subquadratic growth to (3.1), *and v is locally uniformly $\mathcal{C}^{1,\alpha}$. Moreover, there exists $C > 0$ such that for all $r \geq 1$,*

$$\sup_{x\in B_r} |v(x)| \leq C\gamma_r. \tag{3.7}$$

(b) *Assume that there are operators F_λ satisfying* (a)–(c) *uniformly with the above $\gamma_r, \underline{\gamma_r}$ for $\lambda \in (0,1)$, such that $F_\lambda \to F$ as $\lambda \to 0^+$ locally uniformly in all the variables. Then for each $\lambda \in (0,1)$, there exists a unique solution v_λ satisfying* (3.7) *to*

$$F_\lambda(\nabla^2 v_\lambda, \nabla v_\lambda, v_\lambda, x) = 0 \quad \text{in } \mathbb{R}^d, \tag{3.8}$$

and v_λ is locally uniformly $\mathcal{C}^{1,\alpha}$. Moreover, we have $v_\lambda \to v$ locally uniformly as $\lambda \to 0^+$.

(c) *If F (or F_λ) is concave in X, then v (or v_λ) is locally uniformly $\mathcal{C}^{2,\alpha}$.*

Proof. (i) With the comparison principle (Lemma 3.5), we only need to produce a supersolution and a subsolution that have subquadratic growth at infinity, and invoke Perron's method.

By (*a*)–(*c*) and (3.6), there exists a constant $C > 0$ such that for any $x \in \mathbb{R}^d$, $(X,p) \in \mathcal{S}^d \times \mathbb{R}^d$, and $s \geq 0$, if $r := |x| \geq 1$, then

$$F(X,p,s,x) \geq \rho s - \gamma_r(1 + |p|/r) - C(1 + |X|), \tag{3.9}$$

and if $r \in [0,1)$, then

$$F(X,p,s,x) \geq \rho s - C(1 + |p| + |X|). \tag{3.10}$$

Let $\phi \in \mathcal{C}^2([0,\infty))$ be a regularization of $r \to \gamma_r$ such that

$$\phi'(0) = \phi''(0) = 0,\ \phi'(\cdot) \geq 0,\ \phi(r) = \gamma_r \text{ for } r \geq 1, \text{ and } \limsup_{r>0} \phi'(r)/r < \infty. \tag{3.11}$$

Define $\bar{v}(x) := C_1 + C_2\phi(|x|)$ for some $C_1, C_2 > 0$ to be determined. For simplicity, below we drop (x) and $(|x|)$ from the notations of $\bar{v}(x)$, $\phi(|x|)$, $\phi'(|x|)$, and $\phi''(|x|)$. For $|x| \geq 1$, we have from (3.9) that $F(\nabla^2\bar{v}, \nabla\bar{v}, \bar{v}, x) \geq \rho(C_1 + C_2\phi) - \phi(1 + C_2\phi'/|x|) - C(1 + C_2|\phi''|)$. It follows from (3.5) that $\phi'(r)/r + \phi(r)^{-1}(|\phi''(r)| + \phi'(r)) \to 0$ as $r \to \infty$. Therefore by picking C_2 and then C_1 to be sufficiently large, we obtain $F(\nabla^2\bar{v}, \nabla\bar{v}, \bar{v}, x) \geq 0$. This inequality holds the same when $|x| < 1$ by (3.10) and (3.11). Similarly, one can show that $\underline{v} := -\bar{v}$ is a subsolution. Clearly both $\bar{v}$ and $\underline{v}$ have at most subquadratic growth. Thus, by Perron's method and Lemma 3.5 (note that by (3.5), γ_r satisfies (3.2)), we obtain the unique solution v to (3.1), and $\underline{v} \leq v \leq \bar{v}$ yields (3.7). Finally $v \in \mathcal{C}^{1,\alpha}$ follows from Lemma 3.4.

(ii) The above argument also yields the unique solution v_λ to (3.8) with $v_\lambda \in \mathcal{C}^{1,\alpha}$ satisfying (3.7) for each $\lambda \in (0,1)$. Let v_*, v^* be defined as in Lemma 3.2. Since $F_\lambda \to F$ locally uniformly, Lemma 3.2 yields that v_* and v^* are, respectively, a supersolution and a subsolution to (3.1). As v_* and v^* have at most subquadratic growth, applying Lemma 3.5 yields $v_* \geq v^*$ in $\mathbb{R}^d$. The other direction of the inequality holds trivially by definition; hence $v_* = v^*$ which then equals the unique solution v to (3.1). This shows $v_\lambda \to v$ locally uniformly as $\lambda \to 0^+$.

(iii) This follows readily from Lemma 3.4. ∎

3.3. Rate of convergence. Recall that $|X|$ denotes the spectral norm for $X \in \mathcal{S}^d$. We make the following assumption on the difference between F and F_λ:

(d) There exists a continuous function $\omega_0 : [0,\infty)^4 \to [0,\infty)$ such that for each $\lambda \geq 0$, $\omega_0(\lambda, \cdot, \cdot, \cdot)$ is nondecreasing in all its variables, $\omega_0(0, \cdot, \cdot, \cdot) \equiv 0$, and for each $(X, p, s, x) \in \mathcal{S} \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d$ we have

$$|F_\lambda(X, p, s, x) - F(X, p, s, x)| \leq \omega_0(\lambda, |X|, |p|, |x|).$$

In the remainder of this subsection, we derive a convergence rate of $v_\lambda \to v$ as $\lambda \to 0^+$, assuming that the Lipschitz norms of v_λ and v are not too large at $x \to \infty$. To the best of our knowledge, this error estimate result in the general setting with possibly unbounded solutions in $\mathbb{R}^d$ is new.

THEOREM 3.7. *Let $C_0 \geq 1, \eta \in [0,2)$, F, F_λ satisfy* (a)–(d) *with $\gamma_r = C_0(1 + r^\eta)$, $\underline{\gamma_r} = C_0(1 + r^{\eta-1})$, and v and v_λ be, respectively, the solutions to* (3.1) *and* (3.8). *Suppose for some $\alpha \geq 0$, we have for each $r \geq 1$,*

$$|\nabla v(\cdot)| + |\nabla v_\lambda(\cdot)| \leq C_0 r^\alpha \quad \text{in } B_r. \tag{3.12}$$

Then there exist $A, C > 0$ such that for all $\lambda \in (0,1)$ and $r \geq 1$, we have

$$\sup_{x \in B_r} |v_\lambda(x) - v(x)| \leq \rho^{-1}\omega_0(\lambda, R^{c_1}, R^{c_2}, R^{c_3}) + CR^{-c_4},$$

where $R := Ar$, $\varepsilon := (2-\eta)/2$, $c_2 := 1+\varepsilon$, $c_3 := \max\{\alpha(1+\varepsilon), 1\}$, $c_4 := 1 + \min\{(1-\eta)(1+\varepsilon), 0\}$, and $c_1 := (2\alpha + 2\eta)(1+\varepsilon) + c_4$.

Proof. We will only show that v cannot be too much larger than v_λ for $\lambda \in (0,1)$ in B_r; the proof for the other direction is almost identical. From the assumption

and Theorem 3.6, there is $C_1 \geq C_0$ such that for all $r \geq 1$, we have $\gamma_r \leq C_1 r^\eta$, $\underline{\gamma_r} \leq C_1(1 + r^{\eta-1})$, and

$$|v(\cdot)| + |v_\lambda(\cdot)| \leq C_1 r^\eta \quad \text{in } B_r. \tag{3.13}$$

Then after writing $\delta_r := \sup_{x \in B_r}(v(x) - v_\lambda(x))$ for some $r \geq 1$, (3.13) yields $\delta_r \leq C_1 r^\eta$.

Let $R_1 := Ar$ for some $A \geq 1$, and $R_2 := R_1^{1+\varepsilon}$ with $\varepsilon = \frac{2-\eta}{2} \in (0, 1]$. We consider a radially symmetric and radially nondecreasing function $\phi : \mathbb{R}^d \to [0, \infty)$ such that

$$\phi(\cdot) \equiv 0 \text{ on } B_r, \quad \phi(\cdot) \geq C_1 R_2^\eta \text{ on } \partial B_{R_2}, \tag{3.14}$$

and for some $C = C(d)$,

$$|\nabla \phi(x)| \leq C(1 + \phi(x))/R_1, \quad |\nabla^2 \phi(x)| \leq C(1 + \phi(x))/(R_1 r) \tag{3.15}$$

for all $x \in B_{R_2}$. A regularization of the map $x \to \exp(\max\{0, x - r\}/R_1) - 1$ will do if A is large enough depending only on η, C_1. With one fixed A, below we prove a finer bound of δ_r for all r large enough and $\lambda \in (0, 1)$.

Due to (3.13) and (3.14), there exists $x_0 \in B_{R_2}$ such that

$$v(x_0) - v_\lambda(x_0) - 2\phi(x_0) = \sup_{x \in \mathbb{R}^d} (v(x) - v_\lambda(x) - 2\phi(x)) =: \delta' \geq \delta_r. \tag{3.16}$$

Similarly, for any $\beta \geq 1$, we can find $x_1, y_1 \in B_{R_2}$ such that

$$\begin{aligned} &v(x_1) - v_\lambda(y_1) - \phi(x_1) - \phi(y_1) - \beta|x_1 - y_1|^2 \\ &= \sup_{x,y \in \mathbb{R}^d} \left(v(x) - v_\lambda(y) - \phi(x) - \phi(y) - \beta|x - y|^2\right) \geq v(x_0) - v_\lambda(x_0) - 2\phi(x_0) = \delta'. \end{aligned} \tag{3.17}$$

If $\phi(x_1) \leq \phi(y_1)$, noting $|v_\lambda(x_1) - v_\lambda(y_1)| \leq C_0 R_2^\alpha |x_1 - y_1|$ in view of (3.12), we conclude from (3.16) and (3.17) that

$$\begin{aligned} \delta' &\leq v(x_1) - v_\lambda(x_1) - 2\phi(x_1) + C_0 R_2^\alpha |x_1 - y_1| - \beta|x_1 - y_1|^2 \\ &\leq \delta' + C_0 R_2^\alpha |x_1 - y_1| - \beta|x_1 - y_1|^2, \end{aligned}$$

which yields

$$|x_1 - y_1| \leq C_0 R_2^\alpha / \beta. \tag{3.18}$$

This estimate still holds if $\phi(x_1) \geq \phi(y_1)$ by the same argument. Let us write $C_\phi := \phi(x_1) + \phi(y_1)$. It follows from (3.17) that

$$v(x_1) - v_\lambda(y_1) \geq C_\phi + \delta'. \tag{3.19}$$

Since $(1 + \varepsilon)\eta \leq 2$, (3.19) and (3.13) yield $C_\phi \leq C_1 R_2^\eta \leq C_1 R_1^2$.

Now we proceed by making use of (3.17). Since v, v_λ are solutions to (3.1) and (3.8), respectively, the Crandall–Ishii lemma [9, Theorem 3.2] yields that there are matrices $X, Y \in \mathcal{S}^d$ satisfying the following:

$$-(2\beta + |J|)I \leq \begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} \leq J + \frac{1}{2\beta} J^2 \quad \text{with } J := 2\beta \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}, \tag{3.20}$$

and

$$F(X+\nabla^2\phi(x_1),p_1,v(x_1),x_1)\le 0\le F_\lambda(Y-\nabla^2\phi(y_1),q_1,v_\lambda(y_1),y_1), \tag{3.21}$$

where $p_1:=2\beta(x_1-y_1)+\nabla\phi(x_1)$, $\quad q_1:=2\beta(x_1-y_1)-\nabla\phi(y_1)$. Using (c) and (3.21) gives $\rho(v(x_1)-v_\lambda(y_1))\le F_\lambda(Y-\nabla^2\phi(y_1),q_1,v(x_1),y_1)-F(X+\nabla^2\phi(x_1),p_1,v(x_1),x_1)$. Writing $Y':=Y-\nabla^2\phi(y_1)$ and $Z:=X-Y+\nabla^2\phi(x_1)+\nabla^2\phi(y_1)$, we conclude from (a), (b), (d), and $x_1,y_1\in B_{R_2}$ that

$$\begin{aligned}\rho(v(x_1)-v_\lambda(y_1))&\le\omega_0(\lambda,|Y'|,|y_1|,|q_1|)+C_1R_2^\eta|x_1-y_1|(1+|p_1|+|q_1|+|Y'|)\\&\quad+C_1(1+R_2^{\eta-1})|p_1-q_1|+\Lambda_2\mathrm{Tr}(Z)1_{Z\ge0}+\Lambda_1\mathrm{Tr}(Z)1_{Z\le0}.\end{aligned}\tag{3.22}$$

Then we apply (3.13), (3.15), (3.18), and $C_\phi\le C_1R_1^2$ to obtain

$$\begin{aligned}|q_1|&\le C(R_2^\alpha+C_\phi R_1^{-1})\le C(R_2^\alpha+R_1),\\|x_1-y_1|(1+|p_1|+|q_1|+|Y'|)&\le C(R_2^{2\alpha}+C_\phi R_2^\alpha R_1^{-1}+R_2^\alpha|Y'|)/\beta\\&\le C(R_2^{2\alpha}+C_\phi R_2^\alpha R_1^{-1}+R_2^\alpha|Y|)/\beta,\\(1+R_2^{\eta-1})|p_1-q_1|&\le C(1+R_2^{\eta-1})(1+C_\phi)R_1^{-1}\le C(1+C_\phi)R_1^{-c_4},\end{aligned}$$

where $c_4:=1+\min\{(1-\eta)(1+\varepsilon),0\}\in(0,1]$ by $\varepsilon=\frac{2-\eta}{2}$, and $C=C(C_0,C_1)>0$.

Notice that $X\le Y$, and $-6\beta I\le Y\le 6\beta I$ by (3.20). Therefore, (3.15) implies for some $C=C(\Lambda_2)>0$, $\Lambda_2\mathrm{Tr}(Z)1_{Z\ge0}+\Lambda_1\mathrm{Tr}(Z)1_{Z\le0}\le-\Lambda_1\mathrm{Tr}(Y-X)+C(1+C_\phi)R_1^{-1}$. Moreover, it follows from $\beta\ge1$, $R_1=Ar$, and $C_\phi\le CR_1^2$ that for some $C=C(A)>0$, $|Y'|\le|Y|+CC_\phi(R_1r)^{-1}\le C\beta$. Plugging the above estimates into (3.22) shows

$$\begin{aligned}\rho(v(x_1)-v_\lambda(y_1))&\le\omega_0(\lambda,C\beta,R_2,C(R_2^\alpha+R_1))-\Lambda_1\mathrm{Tr}(Y-X)+CR_2^{\alpha+\eta}|Y|/\beta\\&\quad+C(R_2^{2\alpha+\eta}/\beta+R_1^{-c_4})+CC_\phi(R_2^{\alpha+\eta}R_1^{-1}/\beta+R_1^{-c_4}).\end{aligned}\tag{3.23}$$

Notice that by [18, Lemma 3.1] and (3.20), there is $C=C(d)>0$ such that $|X|+|Y|\le C(\mathrm{Tr}(Y-X)+\beta^{\frac12}\mathrm{Tr}(Y-X))$. Therefore, if $2CR_2^{\alpha+\eta}\le\Lambda_1\beta$, we obtain

$$\begin{aligned}CR_2^{\alpha+\eta}|Y|/\beta&\le\Lambda_1\mathrm{Tr}(Y-X)/2+CR_2^{\alpha+\eta}\mathrm{Tr}(Y-X)^{\frac12}\beta^{-\frac12}\\&\le\Lambda_1\mathrm{Tr}(Y-X)+CR_2^{2\alpha+2\eta}/\beta.\end{aligned}$$

Thus, it follows from (3.23) that

$$\begin{aligned}\rho(v(x_1)-v_\lambda(y_1))&\le\omega_0(\lambda,C\beta,R_2,C(R_2^\alpha+R_1))\\&\quad+C(R_2^{2\alpha+2\eta}/\beta+R_1^{-c_4})+CC_\phi(R_2^{\alpha+\eta}R_1^{-1}/\beta+R_1^{-c_4}).\end{aligned}\tag{3.24}$$

Now we pick $\beta:=R_1^{c_1}$ with $c_1:=(1+\varepsilon)(2\alpha+2\eta)+c_4$. Then $\Lambda_1\beta=R_1^{c_1}\ge2CR_1^{(1+\varepsilon)(\alpha+\eta)}=2CR_2^{\alpha+\eta}$ holds when $r\ge1$ (since $R_1=Ar$) is large enough. By (3.24), there exist $C,C'>0$ depending only on C_0,C_1, and η such that $\rho(v(x_1)-v_\lambda(y_1))\le\omega_0\left(\lambda,CR_1^{c_1},R_1^{1+\varepsilon},C(R_1^{\alpha(1+\varepsilon)}+R_1)\right)+CR_1^{-c_4}+C'C_\phi R_1^{-c_4}$. Recall (3.19). Upon further assuming $Ar=R_1\ge(C'/\rho)^{1/c_4}$, we have

$$\rho\delta_r\le\rho\delta'\le\omega_0\left(\lambda,CR_1^{c_1},R_1^{1+\varepsilon},C(R_1^{\alpha(1+\varepsilon)}+R_1)\right)+CR_1^{-c_4}.$$

This leads to the desired conclusion with A replaced by CA, where $A,C>0$ depend only on d,η,C_0,C_1,ρ. ∎

3.4. Exploratory HJB equations: Well-posedness and convergence. Now we apply the general PDE results established in the previous subsections to study the well-posedness of the exploratory HJB equation (1.1) for fixed $\lambda > 0$, as well as the convergence of the solution as $\lambda \to 0^+$.

We assume that the control space $\mathcal{U}$ is a nonempty open subset of some Euclidian space $\mathbb{R}^l$, and let $\rho > 0$. Consider the operator associated with the exploratory HJB equation (1.1),

$$F_\lambda(X,p,s,x) := \rho s - \lambda \ln \int_{\mathcal{U}} \exp\left(\frac{1}{\lambda}(h(x,u) + b(x,u)p + \operatorname{Tr}(\sigma(x,u)\sigma(x,u)^T X))\right) du, \tag{3.25}$$

and the operator associated with the classical HJB equation (2.3),

$$F(X,p,s,x) := \rho s - \sup_{u\in\mathcal{U}} \left(h(x,u) + b(x,u)p + \operatorname{Tr}(\sigma(x,u)\sigma(x,u)^T X)\right). \tag{3.26}$$

We also make the following assumptions on the functions $h(\cdot,\cdot), b(\cdot,\cdot), \sigma(\cdot,\cdot)$.

Assumption 3.8. There are positive $\gamma_r, \underline{\gamma_r} \in \mathcal{C}^2(0,\infty)$ satisfying (3.5) and (3.6) such that the following hold:

(i) For each $r \geq 1$, $|h(\cdot,\cdot)|$ is bounded by γ_r in $B_r \times \mathcal{U}$, and $|b(\cdot,\cdot)|$ is bounded by $\underline{\gamma_r}$ in $B_r \times \mathcal{U}$.
(ii) For each $r \geq 1$ and all $u \in \mathcal{U}$, $h(\cdot,u)$, $b(\cdot,u)$, and $\sigma(\cdot,u)$ are uniformly Lipschitz continuous with Lipschitz bound γ_r in B_r.
(iii) There exist $\Lambda_2 > \Lambda_1 > 0$ such that $\Lambda_1 I \leq \sigma(\cdot,\cdot)\sigma(\cdot,\cdot)^T \leq \Lambda_2 I$ in $\mathbb{R}^d \times \mathcal{U}$.
(iv) $h(\cdot,\cdot), b(\cdot,\cdot), \sigma(\cdot,\cdot)$ are locally uniformly continuous in $\mathbb{R}^d \times \mathcal{U}$.
(v) We have

$$\sup_{\lambda\in(0,1)} \left|\lambda \ln \int_{u\in\mathcal{U}} \exp\left(\frac{h(0,u)}{\lambda}\right) du\right| < \infty, \tag{3.27}$$

and the following holds locally uniformly in $(X,p,x) \in \mathcal{S}^d \times \mathbb{R}^d \times \mathbb{R}^d$:

$$\limsup_{N\to\infty} \sup_{\lambda\in(0,1)} \left|\lambda \ln \int_{u\in\mathcal{U}\setminus[-N,N]^l} \exp\left(\frac{1}{\lambda}(h(x,u) + b(x,u)p + \operatorname{Tr}(\sigma(x,u)\sigma(x,u)^T X)\right) du\right| = 0. \tag{3.28}$$

The condition (3.27) is to ensure that F_λ with $\lambda \in (0,1)$ are well-defined, whereas the condition (3.28) is to guarantee $F_\lambda \to F$ locally uniformly as $\lambda \to 0^+$ which is a reasonable requirement. If $\mathcal{U}$ is a bounded set, then assumption (v) holds trivially. Note that Assumption 3.8 rules out the LQ case (i.e., $b(\cdot,\cdot), \sigma(\cdot,\cdot)$ are linear and $h(\cdot,\cdot)$ quadratic), but the corresponding exploratory and classical HJB equations for LQ can both be solved explicitly and the solutions are quadratic functions; see [32]. In other words, the LQ case can be solved separately and specially and hence is not our concern here.

We have the following result by specializing the results in subsections 3.2–3.3 to the operators F_λ, F defined by (3.25)–(3.26).

THEOREM 3.9. *Let F_λ, F be defined by* (3.25)–(3.26) *and Assumption* 3.8 *hold. Then the assumptions (a)–(d) hold uniformly for F_λ, F for all $\lambda \in (0,1)$, with*

$$\omega_0(\lambda, x_1, x_2, x_3) := \sup_{|X|\leq x_1, |p|\leq x_2, |x|\leq x_3} |F_\lambda(X,p,0,x) - F(X,p,0,x)|,$$

and F_λ, F are concave in X. Consequently, the equation $F_\lambda(\nabla^2 v_\lambda, \nabla v_\lambda, v_\lambda, x) = 0$ (resp., $F(\nabla^2 v, \nabla v, v, x) = 0$) has a unique solution v_λ (resp., v) of subquadratic growth. Moreover,

(i) *v_λ, v are locally $\mathcal{C}^{2,\alpha}$ for some $\alpha \in (0,1)$;*
(ii) *there exists $C > 0$ such that $\sup_{B_r} |v(x)| + |v_\lambda(x)| \le C\gamma_r$ for each $r \ge 1$;*
(iii) *$v_\lambda \to v$ locally uniformly as $\lambda \to 0^+$.*

Proof. It is direct to check that Assumption 3.8 implies assumptions (a)–(c). To see (d), note that if $\mathcal{U}$ is a bounded set, $F_\lambda(X,p,s,x) \to F(X,p,s,x)$ locally uniformly in X,p,s,x as $\lambda \to 0^+$ since $h(x,u), b(x,u), \sigma(x,u)$ are locally uniformly continuous in u and uniformly continuous in x. If $\mathcal{U}$ is unbounded, we use (3.28) to get the convergence.

Clearly the operator F is concave in X according to Definition 3.3. Now we show that F_λ is also concave in X. Let us write, for any fixed p, x, $(a_{ij}) = (a_{ij}(u)) := \sigma(x,u)\sigma(x,u)^T$, $g = g(X,u) := h(x,u) + b(x,u)p + \mathrm{Tr}(\sigma(x,u)\sigma(x,u)^T X)$, and $G = G(X,u) := \exp(\lambda^{-1} g(X,u))$. Then $\frac{\partial g(X,u)}{\partial X_{ij}} = a_{ij}$ and $\frac{\partial^2 g(X,u)}{\partial X_{ij}\partial X_{kl}} = 0$. Direct computation yields that for any $N = (N_{ij}) \in \mathcal{S}^d$,

$$
\begin{aligned}
&- \frac{\partial^2 F_1(X,p,s,x)}{\partial X_{ij}\partial X_{kl}} N_{ij} N_{kl} \\
&= \frac{\left(\int_{\mathcal{U}} G\,du\right)\left(\int_{\mathcal{U}} (\sum_{ij} a_{ij} N_{ij})^2 G\,du\right) - \left(\sum_{ij}\left(\int_{\mathcal{U}} a_{ij} N_{ij} G\,du\right)\right)^2}{\lambda\left(\int_{\mathcal{U}} G\,du\right)^2} \ge 0,
\end{aligned}
$$

where the last inequality is due to Hölder's inequality and $G > 0$. Therefore F_λ is concave in X. All the conclusions now follow from Theorem 3.6. ∎

One can derive a convergence rate for $v_\lambda \to v$ as $\lambda \to 0^+$ in the spirit of Theorem 3.7, but we chose not to present it in the above theorem because its expression would be overly complex for the general case. In the next section, we will derive a simple, explicit rate for a special application case–the temperature control problem.

So far we have focused our attention on the HJB equations. The connection to the control problems is stipulated in the following theorem.

THEOREM 3.10. *Consider the exploratory control problem (2.4)–(2.5) with the value function v_λ. Let Assumption 3.8 hold, and assume that the SDE (2.8) is well-posed. Then v_λ is the unique solution of subquadratic growth to the exploratory HJB equation (1.1). Moreover, v_λ is locally $\mathcal{C}^{2,\alpha}$ for some $\alpha \in (0,1)$, and*

$$
v_\lambda \to v \quad \text{locally uniformly as} \lambda \to 0^+,
$$

where v is the value function of the classical control problem (2.1)–(2.2) and the unique solution of subquadratic growth to the classical HJB equation (2.3).

Proof. Under Assumption 3.8, let v'_λ be the unique solution to (1.1). According to Theorem 3.9(ii), v'_λ has polynomial growth. By a standard verification argument, we have $v_\lambda(x) \le v'_\lambda(x)$ for all $x \in \mathbb{R}^d$. Since (2.8) is well-posed, the equality is achieved by the relaxed control $\pi^*_t(\cdot) = \pi^*(\cdot, X^{\lambda,*}_t)$, namely, $v_\lambda \equiv v'_\lambda$. The remainder of the theorem follows readily from Theorem 3.9. ∎

Theorem 3.10 indicates that the exploratory control problem (2.4)–(2.5) converges to the classical stochastic control problem (2.1)–(2.2) as the weight parameter $\lambda \to 0^+$. The technical assumption needed is that the optimally controlled process $(X^{\lambda,*}_t,\ t \ge 0)$

defined by the SDE (2.8) is well-posed. If $\gamma_r = C(1+r)$ for some $C > 0$ in Assumption 3.8, then it is easy to see that $x \to \widetilde{b}(x, \pi^*(\cdot, x))$ is bounded and measurable, and $x \to \widetilde{\sigma}(\pi^*(\cdot, x))$ is bounded, continuous, and strictly elliptic. Classical theory of [28] then implies that (2.8) is well-posed.

4. Application to exploratory temperature control. In this section we apply the general results obtained in the previous section to the exploratory temperature control problem.

4.1. Exploratory temperature control problem. To design an endogenous temperature control for SA, [15] first consider the following stochastic control problem:

$$\begin{aligned} v(x) :=& \inf \mathbb{E}\bigg[\int_0^\infty e^{-\rho t} f(X_t) dt\bigg], \\ &\text{subject to (1.2) where} \\ &\{\beta_t,\, t \geq 0\} \text{ is adapted, and } \beta_t \in \mathcal{U} \text{ a.e. } t \geq 0, \text{a.s.} \end{aligned} \tag{4.1}$$

Here, the temperature process $(\beta_t,\, t \geq 0)$ is taken as the control. Following [15], we take the control space $\mathcal{U} = [a, 1]$ for a fixed $a \in (0, 1)$ throughout this section.

By setting $\mathcal{U} = [a, 1]$, $h(x, u) = f(x)$, $b(x, u) = -\nabla f(x)$, $\sigma(x, u) = \sqrt{2u}$, and substituting for "sup" with "inf" in (2.3), we obtain the classical HJB equation of the temperature control problem (4.1):

$$-\rho v(x) + f(x) - \nabla f(x) \cdot \nabla v(x) + \inf_{\beta \in [a,1]} \big[\beta \mathrm{Tr}(\nabla^2 v(x))\big] = 0. \tag{4.2}$$

It is then easily seen from the verification theorem that an optimal feedback control has the bang-bang form: $\beta^* = 1$ if $\mathrm{Tr}(\nabla^2 v(x)) < 0$, and $\beta^* = a$ if $\mathrm{Tr}(\nabla^2 v(x)) \geq 0$. Using this temperature control scheme, one should switch between the highest temperature and the lowest one, depending on the sign of $\mathrm{Tr}(\nabla^2 v(x))$. As mentioned in the introduction, there are two disadvantages, one in theory and the other in application, of this bang-bang strategy:

1. Although theoretically optimal, this strategy is too rigid practically to achieve good performance as it only has two actions: $a \to 1$ and $1 \to a$. It is too sensitive to errors which are inevitable in any real world application.
2. The corresponding optimally controlled dynamics is governed by the SDE:

$$dX_t^* = -\nabla f(X_t^*) dt + g(X_t^*) dB_t, \quad X_0^* = x, \tag{4.3}$$

where

$$g(x) := \begin{cases} \sqrt{2a} & \text{if} \mathrm{Tr}(\nabla^2 v(x)) \geq 0, \\ \sqrt{2} & \text{if} \mathrm{Tr}(\nabla^2 v(x)) < 0. \end{cases} \tag{4.4}$$

There is a subtle issue regarding the well-posedness of the SDE (4.3). Note that g is bounded and strictly elliptic. If ∇f is assumed to be bounded, it follows from Exercise 12.4.3 in [28] that (4.3) has a weak solution for all dimension d. However, the uniqueness in distribution may fail since g is *discontinuous* (see, e.g., [26] for an example). According to Exercises 7.3.3 and 7.3.4 in [28], the uniqueness holds for $d = 1, 2$. But it remains unknown whether the uniqueness in distribution is still valid for $d \geq 3$.

To address the first disadvantage above, Gao, Su, and Zhou [15] introduced the exploratory version of (4.1) in order to smooth out the temperature process. This

way, a classical control $(\beta_t,\, t \geq 0)$ is replaced by a relaxed control $\pi = (\pi_t(\cdot),\, t \geq 0)$ over the control space $\mathcal{U} = [a, 1]$, rendering the following exploratory dynamics:

$$dX_t^\pi = -\nabla f(X_t^\pi)dt + \left(\int_{\mathcal{U}} 2u\pi_t(u)du \right)^{\frac{1}{2}} dB_t. \tag{4.5}$$

The exploratory temperature control problem is to solve

$$v_\lambda(x) := \inf_{\pi \in \mathcal{A}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t} f(X_t^\pi)dt - \lambda \int_0^\infty e^{-\rho t} \int_{\mathcal{U}} -\pi_t(u) \ln \pi_t(u) du dt \right], \tag{4.6}$$

where $\mathcal{A}(x)$ is the set of admissible controls specified by Definition 2.1.

The corresponding exploratory HJB equation is

$$-\rho v_\lambda(x) + \nabla f(x) \cdot \nabla v_\lambda(x) + f(x) - \lambda \ln \int_a^1 \exp\left(-\frac{\mathrm{Tr}(\nabla^2 v_\lambda(x))}{\lambda} u \right) du = 0, \tag{4.7}$$

with the optimal feedback control $\pi^*(u; x) = \frac{\exp\left(-\lambda^{-1}\mathrm{Tr}(\nabla^2 v_\lambda(x))u\right)}{\int_a^1 \exp(-\lambda^{-1}\mathrm{Tr}(\nabla^2 v_\lambda(x))u)du}$ for $u \in [a, 1]$, which yields the optimally controlled process governed by the SDE:

$$dX_t^{\lambda,*} = -\nabla f(X_t^{\lambda,*})dt + g_\lambda(X_t^{\lambda,*})dB_t, \tag{4.8}$$

where

$$g_\lambda(x) = \sqrt{2 \frac{\int_a^1 u \exp\left(-\frac{\mathrm{Tr}(\nabla^2 v_\lambda(x))}{\lambda} u \right) du}{\int_a^1 \exp\left(-\frac{\mathrm{Tr}(\nabla^2 v_\lambda(x))}{\lambda} u \right) du}}. \tag{4.9}$$

Note that the diffusion coefficient, g_λ, is now *continuous*, and $\sqrt{2a} \leq g_\lambda(\cdot) \leq 2$. If ∇f is assumed to be bounded, it follows from the classical theory of [28] that (4.8) is well-posed. This is in stark contrast with the controlled dynamics (4.3) which is not necessarily well-posed. In summary, the optimal temperature control scheme of this exploratory formulation allows any level of temperature and renders a well-posed state process, thereby remedying *simultaneously* the two aforementioned disadvantages of the classical formulation.

To study (4.7) and the process governed by (4.8), we make the following assumptions on the function f.

Assumption 4.1. The function $f \in \mathcal{C}^2$ satisfies

(i) there exists a constant $C > 0$ such that $|\nabla f(x)| \leq C$ and $|\nabla^2 f(x)| \leq C(1+|x|)$ for all $x \in \mathbb{R}^d$;
(ii) there exist $\chi > 0$ and $R > 0$ such that $|\nabla f(x)|^2 - d|\nabla^2 f(x)|_{\max} \geq \chi$ for $|x| \geq R$.

Note that a combination of (i) and (ii) yields a linear growth of f. These conditions, in fact, guarantee that both the value function v_λ and the optimal state process $X^{\lambda,*}$ have good properties. We will see that (i) alone is sufficient for identifying the value function v_λ as the solution to the HJB equation, and (ii) is essentially a Lyapunov/Poincaré condition which ensures the convergence of $X^{\lambda,*}$ as $\lambda \to 0^+$.

4.2. Analysis of exploratory HJB equation. In this subsection, we apply the results in section 3 to study (4.7). The corresponding operators are

$$F_\lambda(X, p, s, x) := \rho s - \nabla f(x) \cdot p - f(x) + \lambda \ln \int_a^1 \exp\left(-\frac{\mathrm{Tr}X}{\lambda} u \right) du \tag{4.10}$$

and

$$F(X,p,s,x) := \rho s - \nabla f(x)\cdot p - f(x) - (a1_{\mathrm{Tr}X>0} + 1_{\mathrm{Tr}X<0})\mathrm{Tr}X. \tag{4.11}$$

Specializing Assumption 3.8 to $\mathcal{U} = [a,1]$, $h(x,u) = f(x)$, $b(x,u) = -\nabla f(x)$, and $\sigma(x,u) = \sqrt{2u}$ leads to the following assumption on f.

Assumption 4.2. Assume that $f \in \mathcal{C}^2(\mathbb{R}^d)$, and for each $r \geq 1$,

$$\sup_{|x|<r}(|f(x)| + |\nabla^2 f(x)|) \leq \gamma_r \quad \text{and} \quad \sup_{|x|<r}|\nabla f(x)| \leq \underline{\gamma_r},$$

where $\gamma_r, \underline{\gamma_r} \in \mathcal{C}^2(0,\infty)$ are positive and satisfy (3.5) and (3.6).

Assumption 4.2 basically requires a subquadratic growth on f and a sublinear growth on $|\nabla f|$. It is more general than Assumption 4.1(i). In particular, it recovers Assumption 4.1(i) when $\gamma_r = C(1+r)$.

The following result is an easy corollary of Theorem 3.9.

COROLLARY 4.3. *Let F, F_λ be defined by* (4.10)–(4.11), *and Assumption* 4.2 *hold. Then*

(i) *there exists a unique solution v of subquadratic growth to the equation $F(\nabla^2 v, \nabla v, v, x) = 0$, and v is locally uniformly $\mathcal{C}^{2,\alpha}$;*
(ii) *for each $\lambda > 0$, there exists a unique solution v_λ of subquadratic growth to the equation $F_\lambda(\nabla^2 v_\lambda, \nabla v_\lambda, v_\lambda, x) = 0$, and v_λ is locally uniformly $\mathcal{C}^{2,\alpha}$;*
(iii) *there exists $C \geq 1$ such that for all $r \geq 1$,*

$$\sup_{\lambda\in(0,1)}\sup_{x\in B_r}(|v(x)| + |v_\lambda(x)|) \leq C(1+\gamma_r),$$

and, moreover, $v_\lambda \to v$ locally uniformly as $\lambda \to 0^+$.

Next we apply Theorem 3.7 to derive an explicit rate of convergence for $v_\lambda \to v$ as $\lambda \to 0^+$ by assuming that Assumption 4.2 holds with the choice of $\gamma_r = C(1+r^\eta)$ for some $\eta \in [0,2)$.

LEMMA 4.4. *Let Assumption* 4.2 *hold with $\gamma_r = C(1+r^\eta)$ for some $\eta \in [0,2)$. Then*

(i) *F and F_λ satisfy the assumptions* (a)–(c) *with $\gamma_r = C(1+r^\eta)$, and $\underline{\gamma_r} = C(1+r^{\eta-1})$;*
(ii) *the assumption* (*d*) *holds with*

$$\omega_0(\lambda, x_1, x_2, x_3) := \omega_0(\lambda, x_1) = C\lambda + \lambda\ln(dx_1/\lambda)1_{dx_1>\lambda}, \tag{4.12}$$

where d is the dimesion of the state space.

Proof. The proof of (i) is the same as the one of Theorem 3.9, in which the expression of $\underline{\gamma_r}$ follows from (3.6). The proof of (ii) follows from direct computations, and we will prove (4.12) for the case when $z := \mathrm{Tr}X/\lambda > 0$, the other case being similar. Notice that

$$A_\lambda := F_\lambda(X,p,s,x) - F(X,p,s,x) = \lambda\ln\left[z^{-1}\left(1 - e^{-z(1-a)}\right)\right].$$

If $z \geq 1$ we have $z^{-1}\left(1-e^{-z(1-a)}\right) \in \left[z^{-1}(1-e^{-1+a}), z^{-1}\right]$, and if $z \in (0,1)$ we have $z^{-1}\left(1-e^{-z(1-a)}\right) \in [1-e^{-1+a}, 1-a]$. Therefore, $|A_\lambda| \leq C\lambda + \lambda\ln(z)1_{z>1}$, and the conclusion follows since $d|X| \geq |\mathrm{Tr}X|$. ∎

In the following lemma, we present a pointwise bound of $|\nabla v|$ and $|\nabla v_\lambda|$.

LEMMA 4.5. *Let Assumption* 4.2 *hold with* $\gamma_r = C(1 + r^\eta)$ *for some* $\eta \in [0, 2)$. *Then there exists* $C > 0$ *such that for any* $r \geq 1$ *we have*

$$\sup_{\lambda\in(0,1)} \sup_{x\in B_r} (|\nabla v(x)| + |\nabla v_\lambda(x)|) \leq Cr^\alpha, \quad \text{where } \alpha := \max\{2\eta - 1, \eta\}.$$

Proof. We will only prove for v, and that for v_λ is identical because $F_\lambda, \lambda > 0$, have uniformly elliptic second order terms, while the lower order terms are the same as F.

Fix $r \geq 1$, and let $u(x) := r^{-\eta} v(r^{-\gamma} x)$ with $\gamma := \max\{\eta - 1, 0\}$. According to Corollary 4.3, u is uniformly bounded in $B_{2r^{1+\gamma}}$, and it satisfies $\rho' u - b(x) \cdot \nabla u - c(x) - (a 1_{\Delta u>0} + 1_{\Delta u<0})\Delta u = 0$, where $\rho' := \rho r^{-2\gamma}$, $b(x) := r^{-\gamma}(\nabla f)(r^{-\gamma} x)$, and $c(x) := r^{-2\gamma-\eta} f(r^{-\gamma} x)$. Thus, by the assumption of the lemma and $\gamma \geq \eta - 1$, we have for some $C > 0$,

$$\sup_{r\geq 1} \sup_{x\in B_{2r^{1+\gamma}}} (|b(x)| + |c(x)|) \leq C.$$

This allows us to apply Theorem 2.1 in [22] (see also Theorem 2.1 in [30]) to conclude that $\sup_{x\in B_{r^{1+\gamma}}} |\nabla u(x)| \leq C$ for some C independent of r, completing the proof. ∎

Finally, we state the convergence rate result, the proof of which follows from Theorem 3.7, Lemma 4.4, and Lemma 4.5.

THEOREM 4.6. *Let* F, F_λ *be defined by* (4.10)–(4.11), *and Assumption* 4.2 *hold with* $\gamma_r = C(1 + r^\eta)$ *for some* $\eta \in [0, 2)$. *Also let* v_λ *(resp., v) be the unique solution of subquadratic growth to the equation*

$$F_\lambda(\nabla^2 v_\lambda, \nabla v_\lambda, v_\lambda, x) = 0 \quad (\text{resp., } F(\nabla^2 v, \nabla v, v, x) = 0).$$

Then there exists $C > 0$ *such that for all* $\lambda \in (0, 1)$ *and* $r \geq 1$ *we have*

$$\sup_{x\in B_r} |v_\lambda(x) - v(x)| \leq C\lambda + C\lambda \ln(r/\lambda) + Cr^{-c}$$

with $c := 1 + \min\{(1 - \eta)(4 - \eta)/2, 0\}$.

Combining Theorems 3.10 and 4.6, we get the following result characterizing the value function of the exploratory temperature control problem and its convergence.

COROLLARY 4.7. *Consider the exploratory temperature control problem* (4.5)–(4.6) *with value function* v_λ. *Let Assumption* 4.1(i) *hold. Then* v_λ *is the unique solution of subquadratic growth to the exploratory HJB equation* (4.7). *Moreover,* v_λ *is locally* $\mathcal{C}^{2,\alpha}$ *for some* $\alpha \in (0, 1)$, *and there exists* $C > 0$ *such that for all* $\lambda \in (0, 1)$ *and* $r \geq 1$,

$$\sup_{x\in B_r} |v_\lambda(x) - v(x)| \leq C\lambda + C\lambda \ln(r/\lambda) + Cr^{-1}, \tag{4.13}$$

where v *is the unique solution of subquadratic growth to the classical HJB equation* (4.2).

Because the constant $C > 0$ in (4.13) is independent of $\lambda \in (0, 1)$ and $r \geq 1$, we can minimize the right-hand side of (4.13) with respect to r to get $r_{\min} = \lambda^{-1} > 1$. With $r_{\min}$, (4.13) reduces to

$$\sup_{x\in B_{1/\lambda}} |v_\lambda(x) - v(x)| \le 2C\lambda + 2C\lambda\ln(1/\lambda). \tag{4.14}$$

Note that for many real world optimization problems, one can (and probably *should*) restrict herself to a bounded set—however large it might be—containing all the "important" states. Thus when λ is sufficiently small, the ball of radius $1/\lambda$ contains these states of interest, and the leading term on the right-hand side of (4.14) is $\lambda\ln(1/\lambda)$. Therefore, the estimate (4.14) essentially stipulates that v_λ converges to v at the rate of $\lambda\ln(1/\lambda)$ as $\lambda \to 0^+$.

4.3. Optimally controlled state process. In this subsection we consider the long time behavior of the optimal state process (4.8) of the exploratory temperature control problem.

We start by recalling some basics in stochastic stability. Consider the general diffusion process $X = (X_t,\, t \ge 0)$ in $\mathbb{R}^d$ of form

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \quad X_0 = x, \tag{4.15}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ is the drift, and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the diffusion (or covariance) matrix. Assuming that (4.15) is well-posed, let $\mathcal{L}$ be the infinitesimal generator of the diffusion process X defined by

$$\mathcal{L}\psi(x) = \sum_{i=1}^d b_i(x)\frac{\partial}{\partial x_i}\psi(x) + \frac{1}{2}\sum_{i,j=1}^d \left(\sigma(x)\sigma(x)^T\right)_{ij}\frac{\partial^2}{\partial x_i\partial x_j}\psi(x),$$

and $\mathcal{L}^*$ be the corresponding adjoint operator given by

$$\mathcal{L}^*\psi(x) = -\sum_{i=1}^d \frac{\partial}{\partial x_i}(b_i(x)\psi(x)) + \frac{1}{2}\sum_{i,j=1}^d \frac{\partial^2}{\partial x_i\partial x_j}(\sigma(x)\sigma(x)^T\psi(x))_{ij}, \tag{4.16}$$

where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a suitably smooth test function. The probability density $\rho_t(\cdot)$ of the process X at time t then satisfies the Fokker–Planck equation

$$\frac{\partial\rho_t}{\partial t} = \mathcal{L}^*\rho_t. \tag{4.17}$$

It is not always true that $\rho_t(\cdot)$ converges as $t \to \infty$ to a probability measure. But if b and σ satisfy some growth conditions, it can be shown that as $t \to \infty$, $\rho_t(\cdot)$ converges in total variation distance to $\rho(\cdot)$ which is the stationary distribution (or steady state) of X. It is then easily deduced from (4.17) that ρ is characterized by the equation $\mathcal{L}^*\rho = 0$. For instance, the overdamped Langevin equation with $b(x) = -\nabla f(x)$ and $\sigma(x) = \sqrt{2\beta}\, I$ is time reversible, and the stationary distribution, under some growth condition on f, is the Gibbs measure

$$\mathcal{G}_\beta(dx) := \frac{1}{Z_\beta}\exp\left(-\frac{f(x)}{\beta}\right)dx, \tag{4.18}$$

where $Z_\beta := \int_{\mathbb{R}^d}\exp(-f(x)/\beta)dx$ is the normalizing constant. However, for general b and σ, the stationary distribution $\rho(\cdot)$ may not have a closed-form expression. The standard references for stability of diffusion processes are [11, 24, 23]. We record a result on the ergodicity of diffusion processes.

LEMMA 4.8. *Assume that $b : \mathbb{R}^d \to \mathbb{R}^d$ is bounded, and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is bounded and strictly elliptic, and that there exists $0 < \alpha \le 1$ such that b, σ are locally uniformly α-Hölder continuous, i.e., for each $R > 0$ there is a constant $C_R > 0$ such that*

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| < C_R|x - y|^\alpha \quad \text{for all } \ x, y \in B_R. \tag{4.19}$$

Then (4.15) *is well-posed, i.e., it has a weak solution which is unique in distribution. Assume further that there exist $M_1 > 0$, $M_2 < \infty$, a compact set $C \subset \mathbb{R}^d$, and a function $V : \mathbb{R}^d \to [1, \infty)$ with $V(x) \to \infty$ as $|x| \to \infty$ such that*

$$\mathcal{L}V \le -M_1 + M_2 1_C. \tag{4.20}$$

Then the (unique) distribution of the solution to (4.15) *converges in total variation distance to its unique stationary distribution as $t \to \infty$.*

Proof. The fact that the diffusion process (4.15) is well-posed follows from Theorem 6.2 in [28]. Recall that a Borel set $C \subset \mathbb{R}^d$ is called petite if there exist a distribution q on $\mathbb{R}_+$ and a nonzero Borel measure ν on $\mathbb{R}^d$ such that $\int_0^\infty \mathbb{P}_x(X_t \in A)\, q(dt) \ge \nu(A)$ for all $x \in C$ and all Borel sets $A \subset \mathbb{R}^d$. Under the condition (4.20) with a petite set C, Theorems 2.1 and 2.2 in [31] imply that the diffusion process is positive Harris recurrent, and converges in total variation distance to its unique stationary distribution. Further by Theorem 2.1 in [27], the diffusion process is a Lebesgue irreducible (and T-) process. However, according to Theorem 4.1 in [24], each compact set is petite, which concludes the proof. ∎

The following theorem describes the long time behavior of the optimal state process (4.8) of the exploratory temperature control problem (4.5)–(4.6). Recall that $||\cdot||_{TV}$ denotes the total variation distance between probability measures.

THEOREM 4.9. *Let Assumption* 4.1 *hold. Then we have*

(i) *for each $\lambda > 0$, the process $(X_t^{\lambda,*},\, t \ge 0)$ converges in total variation distance to its unique stationary distribution as $t \to \infty$;*

(ii) *for each $\lambda > 0$, let ρ_λ be the stationary distribution of the process $(X_t^{\lambda,*},\, t \ge 0)$. Fix $\theta > 0$ and $\delta > 0$. Then there exists $c > 0$ such that $\rho_\lambda(\{x : |x - \theta| > \delta\}) > c$ for all $\lambda > 0$. Consequently, $(X_t^{\lambda,*},\, t \ge 0)$ does not converge in probability to any $\theta \in \mathbb{R}^d$ (and in particular to* $\operatorname{argmin} f(x)$*).*

(iii) Let $\mathcal{G}_\beta$, $\beta > 0$, be the Gibbs measure of the form (4.18)*. Then for each $\lambda > 0$, $\rho_\lambda \ne \mathcal{G}_\beta$ for any $\beta > 0$. Moreover, there exists $c > 0$ such that $||\rho_\lambda - \mathcal{G}_\beta||_{TV} > c$ for all $\beta > 0$.*

Proof. (i) Note that $X^{\lambda,*}$ is a diffusion process with $b(x) = -\nabla f(x)$ and $\sigma(x) = g_\lambda(x)I$. It is clear that b is bounded, and σ is bounded and strictly elliptic. By Assumption 4.1(ii), $|\nabla^2 f|$ is bounded, and thus $b = -\nabla f$ satisfies the Hölder condition (4.19). By Corollary 4.3, v_λ is locally $\mathcal{C}^2$. It follows that g_λ is locally Hölder continuous, and so is $\sigma = g_\lambda I$. It is easy to see that

$$\mathcal{L}f(x) = -|\nabla f(x)|^2 + \frac{1}{2}\sum_{i=1}^d g_\lambda^2(x)\frac{\partial^2 f}{\partial x_i^2}(x) \le -|\nabla f(x)|^2 + d|\nabla^2 f(x)|_{\max}.$$

By Assumption 4.1(ii), the condition (4.20) is satisfied with $M_1 = \chi$ and $M_2 = \sup_{x\in B_R} \mathcal{L}f(x)$. It suffices to apply Lemma 4.8 to conclude.

(ii) This follows from the fact that g_λ is bounded away from 0. We argue by contradiction that $\inf_{\lambda>0} \rho_\lambda(\{x : |x - \theta| > \delta\}) = 0$. Then for $\varepsilon > 0$, there exists $\lambda > 0$

such that $\rho_\lambda(\{x : |x-\theta| > \delta\}) < \varepsilon$. By part (i), $(X_t^{\lambda,*}, t \geq 0)$ converges in total variation distance to ρ_λ. So for t sufficiently large, we have

$$\mathbb{P}(X_t^{\lambda,*} > \theta + \delta) < 2\varepsilon. \tag{4.21}$$

On the other hand, $b = -\nabla f$ and $\sigma = g_\lambda I$ are Hölder continuous, and $\sigma\sigma^T \geq 2aI$ with $2a$ independent of λ. By Aronson's comparison theorem (see [2]),

$$\mathbb{P}(X_t^{\lambda,*} > \theta + \delta) \geq C\mathbb{P}(cB_t > \theta + \delta), \tag{4.22}$$

where $c, C > 0$ are constants independent of t and λ. By taking $\varepsilon > 0$ to be arbitrarily small, the estimates (4.21) and (4.22) lead to a contradiction.

(iii) We first prove that $\rho_\lambda \neq \mathcal{G}_\beta$ for any $\beta > 0$. We argue by contradiction that $\rho_\lambda = \mathcal{G}_\beta$ for some $\beta > 0$. Recall from (4.16) that the adjoint operator of the optimal controlled process is $\mathcal{L}^*\psi(x) = -\sum_{i=1}^d \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_i}(x)\psi(x)\right) + \frac{1}{2}\sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}(g_\lambda(x)\psi(x))$ for $\psi : \mathbb{R}^d \to \mathbb{R}$. Since $\mathcal{L}^*\rho_\lambda = 0$, we get

$$-\sum_{i=1}^d \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_i}(x)\rho_\lambda(x)\right) + \frac{1}{2}\sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}(g_\lambda(x)\rho_\lambda(x)) = 0. \tag{4.23}$$

On the other hand, $\rho_\lambda = \mathcal{G}_\beta$ is the stationary distribution of the overdamped Langevin equation $dX_t = -\nabla f(X_t)dt + \sqrt{2\beta}dB_t$; so it satisifies

$$-\sum_{i=1}^d \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_i}(x)\rho_\lambda(x)\right) + \frac{\beta}{2}\sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}\rho_\lambda(x) = 0. \tag{4.24}$$

Comparing (4.23) and (4.24) yields $\Delta(g_\lambda\rho_\lambda - \beta\rho_\lambda) = 0$, i.e. $g_\lambda\rho_\lambda - \beta\rho_\lambda$ is a harmonic function. By Assumption 4.1(ii), $f(x) \to +\infty$ as $|x| \to \infty$. Thus, $g_\lambda\rho_\lambda - \beta\rho_\lambda \to 0$ as $|x| \to \infty$. According to Liouville's theorem, any bounded harmonic function is constant (see, e.g., Theorem 8, Chapter 2 in [12]). So $g_\lambda\rho_\lambda - \beta\rho_\lambda \equiv 0$, and hence $g_\lambda \equiv \beta$. Injecting this into (4.9), we see that v_λ only depends on a, β, and λ. This contradicts the HJB equation (4.7) where v_λ also depends on f.

Now we prove that ρ_λ is bounded away from *any* Gibbs measure $\mathcal{G}_\beta$. We argue by contradiction that $\inf_{\beta>0}||\rho_\lambda - \mathcal{G}_\beta||_{TV} = 0$. Then there exists a sequence $\{\beta_n\}_{n\geq 1}$ such that $||\rho_\lambda - \mathcal{G}_{\beta_n}||_{TV} \to 0$ as $n \to \infty$. This is impossible if $\lim_{n\to\infty}\beta_n = \infty$, since $\mathcal{G}_\beta$ does not converge to a probability measure as $\beta \to \infty$. Thus, we can extract a convergent subsequence $\{\beta'_n\}_{n\geq 1}$ from $\{\beta_n\}_{n\geq 1}$. If $\lim_{n\to\infty}\beta'_n = \beta' > 0$, this implies that $\rho_\lambda = \mathcal{G}_{\beta'}$ which contradicts the fact that $\rho_\lambda \neq \mathcal{G}_\beta$ for any $\beta > 0$. If $\lim_{n\to\infty}\beta'_n = 0$, then ρ_λ is concentrated on $\operatorname{argmin} f$, whose validity is ruled out by part (ii). ∎

Theorem 4.9 indicates that, with a *fixed* level of exploration, the optimally controlled process $(X_t^{\lambda,*}, t \geq 0)$ does have a stationary distribution. This provides a theoretical justification to the SA algorithm devised by [15] based on discretizing (4.8). The result that this stationary distribution is not a Dirac mass on the minimizer of f is expected theoretically because (4.8) is a genuine diffusion process due to its strict ellipticity. It is indeed *preferred* from an exploration point of view because the essence of exploration is to involve as many states as possible instead of just focusing on the single state of the minimizer, in the same spirit of the classical overdamped Langevin diffusion that converges to the Gibbs measure instead of the Dirac one. The fact

that the stationary distribution of (4.8) is *not* a Gibbs measure is the most intriguing one; it suggests the possibility of a greater variety of target measures—beyond Gibbs measures—when it comes to SA for nonconvex optimization.

To conclude this subsection, we study the stability of stationary distributions of $(X_t^{\lambda,*},\, t \geq 0)$ with different λ's. For a general analysis on the stability of stationary distributions of diffusion processes with different drift and covariance coefficients, see [3, 4, 6]. The idea is to bound the total variation distance between stationary distributions in terms of diffusion parameters. We recall a lemma which is due to [6].

LEMMA 4.10. *Let (b_1, σ_1) and (b_2, σ_2) be pairs of drift and covariance coefficients associated with the diffusion process* (4.15). *For each $k = 1, 2$, assume that b_k is bounded and measurable, and σ_k is bounded, strictly elliptic, and globally Lipschitz. Then the diffusion process associated with (b_k, σ_k) has a unique stationary distribution $\rho_k(dx) = \rho_k(x)dx$. For $1 \leq i \leq d$, let $\phi_1^i := b_1^i - \sum_{j=1}^d \frac{\partial}{\partial x_j}(\sigma_1\sigma_1^T)_{ij}$, $\phi_2^i := b_2^i - \sum_{j=1}^d \frac{\partial}{\partial x_j}(\sigma_2\sigma_2^T)_{ij}$, and $\Phi := \frac{(\sigma_1\sigma_1^T - \sigma_2\sigma_2^T)\nabla\rho_2}{\rho_2} - (\phi_1 - \phi_2)$. Assume further that there exist $\kappa > 0$, $M > 0$ and $R > 0$ such that $b_1(x) \cdot x \leq -M|x|^\kappa$ for $|x| > R$. Then there exists $C > 0$ such that $||\rho_1 - \rho_2||_{TV} \leq C \int_{\mathbb{R}^d} |\Phi(x)|\rho_2(dx)$.*

THEOREM 4.11. *Let Assumption* 4.1 *hold, and assume further that there exist $\kappa > 0$, $M > 0$ and $R > 0$ such that*

$$\nabla f(x) \cdot x \geq M|x|^\kappa \quad \text{for} |x| \geq R, \tag{4.25}$$

and that the solution v_λ to (4.7) *is $\mathcal{C}^3$ with bounded third derivatives. For each $\lambda > 0$, let $\rho_\lambda(dx)$ be the stationary distribution of the optimal state process governed by* (4.8). *Then $\lim_{\lambda' \to \lambda} ||\rho_{\lambda'} - \rho_\lambda||_{TV} = 0$.*

Proof. We apply Lemma 4.10 with $b_1(x) = b_2(x) = -\nabla f(x)$, and $\sigma_1(x) = g_{\lambda'}(x)I$, $\sigma_2(x) = g_\lambda(x)I$. In this case, $\Phi(x) = (g_{\lambda'}(x) - g_\lambda(x))\frac{\nabla\rho_\lambda(x)}{\rho_\lambda(x)} + \nabla(g_{\lambda'} - g_\lambda)(x)$. It is easy to see that $\Phi(x) \to 0$ as $\lambda' \to \lambda$. Since v_λ has bounded third derivatives, we have g_λ is globally Lipschitz. Because $b_2 = -\nabla f$ is bounded and $\sigma_2 = g_\lambda I$ is bounded, Lipschitz, and strict elliptic, it follows from Theorem 3.1.2 in [5] that $\int_{\mathbb{R}^d} \left|\frac{\nabla\rho_\lambda(x)}{\rho_\lambda(x)}\right| \rho_\lambda(dx) \leq \sqrt{\int_{\mathbb{R}^d} \left|\frac{\nabla\rho_\lambda(x)}{\rho_\lambda(x)}\right|^2 \rho_\lambda(dx)} < \infty$. By the dominated convergence theorem, we get $\int_{\mathbb{R}^d} |\Phi(x)|\rho_\lambda(dx) \to 0$ as $\lambda' \to \lambda$. It suffices to apply Lemma 4.10 to conclude. ∎

The assumption (4.25) is a version of the dissipative condition, which is standard in Langevin sampling and optimization. The assumption that $|\nabla f|$ is bounded restricts the range of the dissipative exponent κ to $(0, 1]$. The only technical assumption in Theorem 4.11 is that the solution v_λ to the exploratory HJB equation (4.7) is three times continuously differentiable with bounded third derivatives. It implies that $\nabla^2 v_\lambda$ is continuously differentiable and is globally Lipschitz, which is stronger than the result of Theorem 3.6 that $\nabla^2 v$ is locally Hölder continuous. It is interesting to know whether Assumption 4.1 (possibly with some additional conditions on f) implies the boundedness of third derivatives of the solution to (4.7).

5. Finite time horizon. The exploratory control problem (2.4)–(2.5) is a relaxed control problem in the infinite time horizon, and the associated exploratory HJB equation is, therefore, elliptic. Nevertheless, the previous analysis can be adapted, to the extent it can, to the finite time setting where the HJB equation is parabolic.

We follow the formulation in Zhou [37]. Fix $T > 0$, and consider the stochastic control problem whose value function is

$$v(t,x) = \sup_{u\in\mathcal{A}_0(t,x)} \mathbb{E}\left[\int_t^T h_1(t,X_s^u,u_s)ds + h_2(X_T^u)\Big| X_t^u = x\right], \quad (t,x)\in[0,T]\times\mathbb{R}^d, \tag{5.1}$$

where $h_1 : [0,T]\times\mathbb{R}^d\times\mathcal{U}\to\mathbb{R}$ and $h_2 : \mathbb{R}^d\to\mathbb{R}$ are reward functions, and $\mathcal{A}_0(t,x)$ is the set of admissible classical controls with respect to $X_t^u = x$. The state dynamics is

$$dX_t^u = b(t,X_t^u,u_t)dt + \sigma(t,X_t^u,u_t)dB_t. \tag{5.2}$$

Note here b,σ,h_1 depend on t explicitly.

Denote by ∂_t the partial derivative in t, and by ∇_x and ∇_x^2 the gradient and Hessian in x, respectively. The classical HJB equation associated with the problem (5.1)–(5.2) is

$$\begin{cases} \partial_t v(t,x) + \sup_{u\in\mathcal{U}}\Big[h_1(t,x,u) + b(t,x,u)\cdot\nabla_x v(t,x) \quad 0\le t\le T,\\ \quad +\frac{1}{2}\mathrm{Tr}(\sigma(t,x,u)\sigma(x,u)^T\nabla_x^2 v(t,x))\Big] = 0,\\ v(T,x) = h_2(x). \end{cases} \tag{5.3}$$

It is known that a smooth solution to the HJB equation (5.3) gives the value function (5.1). The optimal control at time t is $u_t^* = u^*(t,X_t^*)$, where $u^* : [0,T]\times\mathbb{R}^d\to\mathcal{U}$ is a deterministic mapping obtained by solving the "$\sup_{u\in\mathcal{U}}$" term in (5.3), and the optimally controlled process, given $X_0^* = x_0$, is governed by

$$dX_t^* = b(t,X_t^*,u^*(t,X_t^*))dt + \sigma(t,X_t^*,u^*(t,X_t^*))dB_t, \quad X_0^* = x_0,$$

provided that it is well-posed.

The exploratory control problem with finite time horizon is for solving an entropy-regularized relaxed control problem whose value function is

$$v_\lambda(t,x) = \sup_{\pi\in\mathcal{A}(t,x)} \mathbb{E}\Big[\int_t^T\Big(\int_{\mathcal{U}} h_1(t,X_s^\pi,u)\pi_s(u)du - \lambda\int_{\mathcal{U}}\pi_s(u)\ln\pi_s(u)du\Big)ds + h_2(X_T^\pi)\Big| X_t^\pi = x\Big], \tag{5.4}$$

where $\mathcal{A}(t,x)$ is the set of distributional control processes defined similarly to the infinite horizon setting, and the exploratory dynamics is

$$dX_t^\pi = \widetilde{b}(t,X_t^\pi,\pi_t)dt + \widetilde{\sigma}(t,X_t^\pi,\pi_t)dB_t \tag{5.5}$$

with $\widetilde{b}(t,x,\pi) := \int_{\mathcal{U}} b(t,x,u)\pi(u)du$ and $\widetilde{\sigma}(t,x,\pi) := (\int_{\mathcal{U}}\sigma(t,x,u)\sigma(t,x,u)^T\pi(u)du)^{\frac{1}{2}}$.

A similar argument as in section 2.2 shows that the optimal feedback control at time t is

$$\pi^*(u,t,x) = \frac{\exp\left(\frac{1}{\lambda}\left[h(t,x,u) + b(t,x,u)\cdot\nabla_x v_\lambda(t,x) + \frac{1}{2}\mathrm{Tr}(\sigma(t,x,u)\sigma(t,x,u)^T\nabla_x^2 v_\lambda(t,x))\right]\right)}{\int_{\mathcal{U}}\exp\left(\frac{1}{\lambda}\left[h(t,x,u) + b(t,x,u)\cdot\nabla_x v_\lambda(t,x) + \frac{1}{2}\mathrm{Tr}(\sigma(t,x,u)\sigma(t,x,u)^T\nabla_x^2 v_\lambda(t,x))\right]\right)du}$$

the exploratory HJB equation is the following nonlinear parabolic PDE:

$$\begin{cases} \partial_t v_\lambda(t,x) + \lambda \ln \int_{\mathcal{U}} \exp\left(\frac{1}{\lambda}\Big[h(t,x,u) + b(t,x,u)\cdot \nabla_x v_\lambda(t,x) \\ \qquad\qquad + \frac{1}{2}\mathrm{Tr}(\sigma(t,x,u)\sigma(t,x,u)^T \nabla_x^2 v_\lambda(t,x))\Big]\right) du = 0, \quad 0 \le t \le T, \\ v_\lambda(T,x) = h_2(x), \end{cases} \tag{5.6}$$

and the optimal state process, given $X_0^{\lambda,*} = x_0$ is governed by

$$dX_t^{\lambda,*} = \widetilde{b}(t, X_t^{\lambda,*}, \pi^*(\cdot,t,X_t^{\lambda,*}))dt + \widetilde{\sigma}(t, X_t^{\lambda,*}, \pi^*(\cdot,t,X_t^{\lambda,*}))dB_t, \quad X_0^{\lambda,*} = x_0, \tag{5.7}$$

provided that it is well-posed.

For general fully nonlinear parabolic PDEs, the solution is only known to be $\mathcal{C}_{t,x}^{\alpha,1+\alpha}$ for some $\alpha \in (0,1)$. We record this fact in the following proposition.

PROPOSITION 5.1. *Let Assumption* 3.8 *hold for* $h_1(\cdot,\cdot,\cdot), b(\cdot,\cdot,\cdot), \sigma(\cdot,\cdot,\cdot)$, *and assume that* $h_2(\cdot)$ *satisfies* $|h_2(\cdot)| \le \gamma_r$ *in* B_r. *Then the HJB equation* (5.6) *(resp.,* (5.3)*) has a unique solution* v_λ *(resp.,* v*) of sub-quadratic growth for* $t \in [0,T]$. *Moreover,*

(i) v_λ, v *are* $\mathcal{C}_{t,x}^{\alpha,1+\alpha}$ *locally uniformly in* $[0,T)\times\mathbb{R}^d$ *for some* $\alpha \in (0,1)$;
(ii) *there exists* $C > 0$ *such that* $\sup_{x\in B_r, t\in[0,T]}(|v(t,x)| + |v_\lambda(t,x)|) \le C\gamma_r$ *for each* $r \ge 1$;
(iii) $v_\lambda \to v$ *locally uniformly as* $\lambda \to 0^+$.

We refer to [34, 35] and [10] for the interior pointwise regularity estimate for fully nonlinear parabolic PDEs.

To identify the value function (5.4) (resp., (5.1)) as the solution to the HJB equation (5.6) (resp., (5.3)), the verification theorem requires that these solutions be $\mathcal{C}_{t,x}^{1,2}$. Since the operators in (5.3) and (5.6) are concave in the sense of Definition 3.3, after further assuming F, F_λ, and h_2 to be sufficiently smooth (see [20, 21]), we know from [21, Theorems 6.4.3 and 6.4.4] that v_λ, v are $\mathcal{C}_{t,x}^{2+\alpha}$ locally uniformly in $[0,T)\times\mathbb{R}^d$. Combining this with Theorem 5.1, we get the following analogue of Theorem 3.10 for the exploratory control problem in a finite time horizon.

THEOREM 5.2. *Consider the exploratory control problem* (5.4)–(5.5) *whose value function is* v_λ. *Let the assumptions in Theorem* 5.1 *hold. Assume that the unique solutions to* (5.3) *and* (5.6) *are locally uniformly* $\mathcal{C}_{t,x}^{1,2}$, *and that the SDE* (5.7) *is well-posed. Then* v_λ *is the unique solution of subquadratic growth to the exploratory HJB equation* (5.6). *Moreover,* v_λ *is locally* $\mathcal{C}_{t,x}^{1,2}$, *and*

$$v_\lambda \to v \quad \text{locally uniformly as } \lambda \to 0^+,$$

where v *is the value function of the classical control problem* (5.1)–(5.2) *and the unique solution of subquadratic growth to the classical HJB equation* (5.3).

We can also obtain the rate of convergence in parallel as in Theorem 3.7, and leave the details to interested readers.

6. Conclusions. In this paper, we study the exploratory HJB equation arising from a continuous-time RL framework—that of the exploratory control—put forth by Wang et al. [32]. We establish the well-posedness and regularity of its solution under general assumptions on the system dynamics parameters. This allows for identifying

the value function of the exploratory control problem in general cases, which goes beyond the LQ setting. We also establish a connection between the exploratory control problem and the classical stochastic control problem by showing that the value function of the former converges to that of the latter as the weight parameter for exploration tends to zero. We then apply our general theory to a special example – the exploratory temperature control problem originally introduced by Gao et al. [15] as a variant of SA. We provide a detailed analysis of the problem, with an explicit rate of convergence derived as the weight parameter vanishes. We also consider the long time behavior of the associated optimally controlled process, and study properties of its stationary distribution. The tools that we develop in this paper encompass stochastic control theory, partial differential equations and probability theory.

Acknowledgments. We thank Yufei Zhang for pointing out the literature [21], which allowed us to develop the results for the exploratory control problem in a finite time horizon.

REFERENCES

[1] S. N. Armstrong and H. V. Tran, *Viscosity solutions of general viscous Hamilton-Jacobi equations*, Math. Ann., 361 (2015), pp. 647–687.

[2] D. G. Aronson, *Bounds for the fundamental solution of a parabolic equation*, Bull. Amer. Math. Soc., 73 (1967), pp. 890–896.

[3] V. I. Bogachev, A. I. Kirillov, and S. V. Shaposhnikov, *The Kantorovich and variation distances between invariant measures of diffusions and nonlinear stationary Fokker-Planck-Kolmogorov equations*, Math. Notes, 96 (2014), pp. 855–863.

[4] V. I. Bogachev, A. I. Kirillov, and S. V. Shaposhnikov, *Distances between stationary distributions of diffusions and the solvability of nonlinear Fokker-Planck-Kolmogorov equations*, Teor Veroyatn. Primen, 62 (2017), pp. 16–43.

[5] V. I. Bogachev, N. V. Krylov, M. Röckner, and S. V. Shaposhnikov, *Fokker-Planck-Kolmogorov Equations*, Math. Surv. Monogr. 207, American Mathematical Society, Providence, RI, 2015.

[6] V. I. Bogachev, M. Röckner, and S. V. Shaposhnikov, *The Poisson equation and estimates for distances between stationary distributions of diffusions*, J. Math. Sci. (N.Y.), 232 (2018), pp. 254–282.

[7] L. A. Caffarelli and X. Cabré, *Fully Nonlinear Elliptic Equations*, Amer. Math. Soc. Colloq. Publ. 43, American Mathematical Society, Providence, RI, 1995.

[8] I. Capuzzo-Dolcetta, F. Leoni, and A. Vitolo, *The Alexandrov-Bakelman-Pucci weak maximum principle for fully nonlinear equations in unbounded domains*, Comm. Partial Differential Equations, 30 (2005), pp. 1863–1881.

[9] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[10] M. G. Crandall, M. Kocan, and A. Świech, *L^p-theory for fully nonlinear uniformly parabolic equations*, Comm. Partial Differential Equations, 25 (2000), pp. 1997–2053.

[11] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley Ser. Probab. Math. Stat., Wiley, Hoboken, NJ, 1986.

[12] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, 2nd ed., American Mathematical Society, Providence, RI, 2010.

[13] D. Firoozi and S. Jaimungal, *Exploratory LQG mean field games with entropy regularization*, Automatica J. IFAC, 139 (2022), 110177.

[14] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd ed., Stoch. Model. Appl. Probab. 25, Springer, New York, 2006.

[15] X. Gao, Z. Q. Xu, and X. Y. Zhou, *State-dependent temperature control for Langevin diffusions*, SIAM J. Control Optim., 60 (2022), pp. 1250–1268.

[16] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss. 224, Springer, Berlin, 1983.

[17] X. Guo, R. Xu, and T. Zariphopoulou, *Entropy regularization for mean field games with learning*, Math. Oper. Res., (2022).

[18] H. Ishii and P.-L. Lions, *Viscosity solutions of fully nonlinear second-order elliptic partial differential equations*, J. Differential Equations, 83 (1990), pp. 26–78.

[19] S. Koike and O. Ley, *Comparison principle for unbounded viscosity solutions of degenerate elliptic PDEs with gradient superlinear terms*, J. Math. Anal. Appl., 381 (2011), pp. 110–120.
[20] N. Krylov, *Boundedly nonhomogeneous elliptic and parabolic equations*, Izv Akad. Nauk. SSSR Ser. Mat., 46 (1982), pp. 487–523.
[21] N. Krylov, *Nonlinear Elliptic and Parabolic Equations of the Second Order*, Reidel Publishing, Norwell, MA, 1987.
[22] Y. Lian, L. Wang, and K. Zhang , *Pointwise Regularity for Fully Nonlinear Elliptic Equations in General Forms*, preprint, https://arxiv.org/abs/2012.00324 (2020).
[23] S. P. Meyn and R. L. Tweedie, *Stability of Markovian processes.* II. *Continuous-time processes and sampled chains*, Adv. Appl. Probab., 25 (1993), pp. 487–517.
[24] S. P. Meyn and R. L. Tweedie, *Stability of Markovian processes.* III. *Foster-Lyapunov criteria for continuous-time processes*, Adv. Appl. Probab., 25 (1993), pp. 518–548.
[25] C. Reisinger and Y. Zhang, *Regularity and stability of feedback relaxed controls*, SIAM J. Control Optim., 59 (2021), pp. 3118–3151.
[26] M. V. Safonov, *Nonuniqueness for second-order elliptic equations with measurable coefficients*, SIAM J. Math. Anal., 30 (1999), pp. 879–895.
[27] O. Stramer and R. L. Tweedie, *Existence and stability of weak solutions to stochastic differential equations with non-smooth coefficients*, Statist. Sinica, 7 (1997), pp. 577–593.
[28] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Grundlehren Math. Wiss. 233, Springer, Berlin, 1979.
[29] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, in Adaptive Computation and Machine Learning, 2nd ed., MIT Press, Cambridge, MA, 2018.
[30] A. Swiech, *$W^{1,p}$-interior estimates for solutions of fully nonlinear, uniformly elliptic equations*, Adv. Differential Equations, 2 (1997), pp. 1005–1027.
[31] W. Tang, *Exponential ergodicity and convergence for generalized reflected Brownian motion*, Queueing Syst., 92 (2019), pp. 83–101.
[32] H. Wang, T. Zariphopoulou, and X. Y. Zhou, *Reinforcement learning in continuous time and space: A stochastic control approach*, J. Mach. Learn. Res., 21 (2020), pp. 1–34.
[33] H. Wang and X. Y. Zhou, *Continuous-time mean-variance portfolio selection: A reinforcement learning framework*, Math. Finance, 30 (2020), pp. 1273–1308.
[34] L. Wang, *On the regularity theory of fully nonlinear parabolic equations*. I, Comm. Pure Appl. Math., 45 (1992), pp. 27–76.
[35] L. Wang, *On the regularity theory of fully nonlinear parabolic equations*. II, Comm. Pure Appl. Math., 45 (1992), pp. 141–178.
[36] J. Yong and X. Y. Zhou, *Stochastic Controls – Hamiltonian Systems and HJB Equations*, Appl. Math. (New York) 43, Springer, New York, 1999.
[37] X. Y. Zhou , *Curse of Optimality, and How Do We Break It*, preprint, SSRN:3845462, 2021.