Global Solutions to Nonconvex Problems by Evolution of Hamilton-Jacobi PDEs

Howard Heaton
Typal Research
Typal LLC

Samy Wu Fung
Dept. of Applied Mathematics and Statistics
Colorado School of Mines

Stanley Osher
Dept. of Mathematics
UCLA

Abstract

Computing tasks may often be posed as optimization problems. The objective functions for real-world scenarios are often nonconvex and/or nondifferentiable. State-of-the-art methods for solving these problems typically only guarantee convergence to local minima. This work presents Hamilton-Jacobi-based Moreau Adaptive Descent (HJ-MAD), a zero-order algorithm with guaranteed convergence to global minima, assuming continuity of the objective function. The core idea is to compute gradients of the Moreau envelope of the objective (which is "piece-wise convex") with adaptive smoothing parameters. Gradients of the Moreau envelope (*i.e.* proximal operators) are approximated via the Hopf-Lax formula for the viscous Hamilton-Jacobi equation. Our numerical examples illustrate global convergence.

1 Introduction

Standard data-oriented tasks such as neural network training, parameter estimation in physical models, and phase recovery may be cast as optimization problems. These problems are often highly non-convex, and practical schemes to solve for global minimizers are scarce. State-of-the-art methods are often either impractical [39], heuristic [36, 44], or only guarantee convergence to local minima [10, 25].

In this work, we introduce a zero-order algorithm with guaranteed convergence to global minima. Our approach minimizes Moreau envelopes of objective functions (*e.g.* see Figure 1). To compute gradients of the Moreau

Figure 1: The iterates $\{x^k\}$ generated by HJ-MAD converge to a global minimizer of f (black) via gradient descent on Moreau envelopes (blue).

envelope, we leverage a connection to Hamilton-Jacobi (HJ) partial differential equations (PDEs) via the Hopf-Lax formula. This yields an explicit solution to the HJ equation from the Cole-Hopf formula, giving the name Hamilton-Jacobi-based Moreau Adaptive Descent (HJ-MAD). These gradients take the form of expectation formulas that can be estimated via sampling.

Contribution Our key contributions for *global* minimization of nonconvex functions are as follows.

▶ Present Moreau Adaptive Descent (MAD), a zero-order method for minimization.

▶ Prove function value convergence by MAD to the *global* minimum value.

▶ Connect MAD to an inviscid Burgers' HJ equation, with adaptive time steps.

▶ Efficiently approximate Moreau envelopes and proximals using viscous HJ equations.

Algorithm 1 Moreau Adaptive Descent (MAD)

1: $\text{MAD}(x^1, \alpha, \tau, t_1, T)$:   ◁ Input parameters
2:    **for** $k = 1, 2, \dots$ ◁ Loop until convergence
3:       $\hat{x}^k \in \text{prox}_{t_k f}(x^k)$ ◁ Compute local minimizer
4:       $g^k = t_k^{-1}(x^k - \hat{x}^k)$ ◁ Estimate envelope gradient
5:       $x^{k+1} = x^k - \alpha_k t_k g^k$ ◁ Gradient update
6:       $t_{k+1} = \text{TimeStep}(t_k, g^k, g^{k-1}; \tau, T)$ ◁ Evolve time via (14)
7:    **return** x^k ◁ Output solution estimate

---

2   Moreau Adaptive Descent

For a continuous and bounded from below function $f: \mathbb{R}^n \to \mathbb{R}$, consider the minimization problem

$$\min_{x \in \mathbb{R}^n} f(x). \tag{1}$$

Given a time $t \in (0, \infty)$, the proximal prox_{tf} and Moreau envelope $u(\cdot, t)$ [2, 34] are defined by

$$\text{prox}_{tf}(x) \triangleq \underset{z \in \mathbb{R}^n}{\text{argmin}} \, f(z) + \frac{1}{2t}\|z - x\|^2 \quad \text{and} \quad u(x, t) \triangleq \min_{z \in \mathbb{R}^n} f(z) + \frac{1}{2t}\|z - x\|^2. \tag{2}$$

These quantities are closely tied as the proximal is the set of minimizers defining the envelope. As shown in Figure 1, the envelope u widens valleys of f and its local minimizers align with local minimizers of f (see Lemma A.4). Under certain assumptions, increasing the time t dissipates the envelope enough that all local minimizers of u are global minimizers of f (see Lemma A.6). Leveraging this fact, we generate a sequence $\{x^k\}$ via gradient descent on the envelope $u(\cdot, t_k)$ while evolving time t_k (forward and backward). For practical purposes (discussed in Section 4), time steps t_k are kept relatively small to ensure updates leverage the local landscape of f when possible. Algorithm 1 presents Moreau Adaptive Descent (MAD) (*n.b.* time stepping is defined below in (14)).

Typical results for zero-order methods assume f is smooth; yet, MAD does not require such regularity.

Assumption 2.1. *The function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous.*

The standard first order necessary condition for optimality is $\nabla f = 0$. Weaker versions of this are used when f is merely continuous and nonconvex. We utilize the following definition [16, Section 2], which generalizes both gradients and the notion of subdifferential common in convex settings.[1]

Definition 2.1. *For a function $f: \mathbb{R}^n \to \mathbb{R}$, the subdifferential of f at $\overline{x}$, denoted by $\partial f(\overline{x})$, is the set of all $v \in \mathbb{R}^n$ satisfying*

$$f(x) \geq f(\overline{x}) + \langle v, x - \overline{x} \rangle + o(\|x - \overline{x}\|), \quad as \; x \to \overline{x}. \tag{3}$$

We next assume the set of global minima is compact and distinguishable from other extrema.

Assumption 2.2. *There is $\gamma > 0$ such that*

   i) the set $\mathcal{S}_\gamma \triangleq \{z \in \mathbb{R}^n : f(z) \leq \inf f + \gamma\}$ is compact;

   ii) if $0 \in \partial f(x)$ and $x \in \mathcal{S}_\gamma$, then x is a global minimizer of f.

Lastly, we provide restrictions on the step size α and initial time t_1, assuming $\eta_- \in (0, 1)$.

Assumption 2.3. *The MAD parameters satisfy:*

   i) step size satisfies $\alpha \in (1 - \sqrt{\eta_-}, 1 + \sqrt{\eta_-})$;

   ii) For a global minimizer $x^\star$ of f, times $t_1, \tau, T \in (0, \infty)$ are chosen such that $t_1 \geq \tau$ and $T \geq t_1 \geq \|x^\star - x^1\|^2/2\gamma$, where γ is as in Assumption 2.2.

Combining our assumptions leads to our main result.

Theorem 2.1. *(*GLOBAL MINIMIZATION*) If Assumptions 2.1, 2.2, and 2.3 hold, then the iteration in Lines 3 to 6 of Algorithm 1 yields convergence to optimal objective values, i.e.*

$$\lim_{k \to \infty} f(x^k) = \min_{x \in \mathbb{R}^n} f(x). \tag{4}$$

Additionally, a subsequence $\{x^{n_k}\} \subseteq \{x^k\}$ converges to a global minimizer of f.

---

[1]The usual notion of subdifferential removes $o(\|x - \overline{x}\|)$, instead requiring the inequality to hold for all $x, \overline{x}$.

Algorithm 2 Hamilton-Jacobi Moreau Adaptive Descent (HJ-MAD)

1: HJ-MAD$(x^1, \alpha, t_1, \tau, T)$:	◁ Input parameters
2: **for** $k = 1, 2, \ldots$	◁ Loop until convergence
3: $\quad g^k \leftarrow \nabla u^\delta(x^k, t_k)$	◁ Generate gradient estimate via (15)
4: $\quad x^{k+1} \leftarrow x^k - \alpha t_k g^k$	◁ Gradient update
5: $\quad t_{k+1} \leftarrow \text{TimeStep}(t_k, g^k, g^{k-1}; \tau, T)$	◁ Evolve time via (14)
6: **return** x^k	◁ Output solution estimate

3 Connections to Hamilton-Jacobi PDEs

When the envelope u is differentiable at x^k, its gradient is precisely g^k (see Lemma A.3), *i.e.*

$$g^k = \frac{x^k - \hat{x}^k}{t_k} = \nabla u(x^k, t_k), \quad \text{where} \quad \hat{x}^k = \text{prox}_{t_k f}(x^k). \tag{5}$$

Although the idea to use gradients of $u(x^k, t_k)$ is simple, computing $\hat{x}^k$ and g^k can be as difficult as the original problem (1). This difficulty can be (approximately) circumvented by using a PDE formulation. The envelope u is a special case of the Hopf-Lax formula [18] for PDEs. It can be shown (*e.g.* see [19, Theorem 3.2]) that u is a (weak) solution to Burgers' Hamilton-Jacobi equation[2]

$$\begin{cases} u_t + \frac{1}{2}\|Du\|^2 = 0 & \text{in } \mathbb{R}^n \times (0, T] \\ u = f & \text{on } \mathbb{R}^n \times \{t = 0\}. \end{cases} \tag{6}$$

The key step in obtaining an explicit expression for each subgradient is to approximate the solution u to (6) by adding a small amount of viscosity via a Laplacian term. Namely, fixing $\delta > 0$, we approximate solutions to (6) via the solution u^δ of the associated viscous Burgers' equation

$$\begin{cases} u_t^\delta + \frac{1}{2}\|Du^\delta\|^2 = \frac{\delta}{2}\Delta u^\delta & \text{in } \mathbb{R}^n \times (0, T] \\ u^\delta = f & \text{on } \mathbb{R}^n \times \{t = 0\}. \end{cases} \tag{7}$$

A key result justifying this approximation is that of Crandall and Lyons [12, Theorem 5.1].

Theorem 3.1. *If* $\delta, T \in (0, \infty)$ *and* f *is bounded and Lipschitz, then there is* $C \in (0, \infty)$ *such that*

$$\sup_{t \in [0,T]} \sup_{x \in \mathbb{R}^n} \left| u(x, t) - u^\delta(x, t) \right| \leq C\sqrt{\delta}. \tag{8}$$

This establishes uniform convergence $u^\delta \to u$ as $\delta \to 0^+$. Although f is not necessarily bounded and Lipschitz continuous on $\mathbb{R}^n$, the sequence $\{x^k\}$ generated by Algorithm 1 is bounded (see Lemma A.7). Thus, the above result still applies since f is bounded and Lipschitz continuous on a compact domain containing $\{x^k\}$. Consequently, for $\delta > 0$ sufficiently small, one is, for all practical purposes, justified in using Algorithm 2 to estimate solutions to (1).

Using the transformation $v^\delta \triangleq \exp(-u^\delta/\delta)$, originally attributed to Cole and Hopf [11, 18], it follows that v^δ solves the heat equation, *i.e.*

$$\begin{cases} v_t^\delta - \frac{\delta}{2}\Delta v^\delta = 0 & \text{in } \mathbb{R}^n \times (0, T] \\ v^\delta = \exp(-f/\delta) & \text{on } \mathbb{R}^n \times \{t = 0\}. \end{cases} \tag{9}$$

This transformation to v^δ is of particular interest since v^δ can be expressed via the convolution formula (*e.g.* see [18] for a derivation)

$$v^\delta(x, t) = \left(\Phi_{\delta t} * \exp(-f/\delta) \right)(x), \tag{10}$$

where $\Phi_{\delta t}$ is the fundamental solution to the heat equation, *i.e.*

$$\Phi_{\delta t}(x) \triangleq \begin{cases} (2\pi \delta t)^{-n/2} \cdot \exp(-\frac{|x|^2}{2\delta t}) & \text{in } \mathbb{R}^n \times (0, \infty) \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

---

[2]We call (6) *Burgers'* since, if $\lim_{|x|\to\infty} |u| = 0$, the original Burgers' PDE is obtained via integration by parts.

Thus, using algebraic manipulations, we recover the viscous Burgers' solution

$$u^\delta(x, t) = -\delta \ln \left(\Phi_{\delta t} * \exp(-f/\delta) \right)(x) \ \text{ in } \mathbb{R}^n \times (0, T]. \tag{12}$$

See Appendix C for an intuitive, informal argument that u^δ expressed by (12) converges pointwise to the envelope u as $\delta \to 0$. Moreover, for δ sufficiently small, if u is differentiable at (x, t), then

$$\nabla u(x, t) \approx \nabla u^\delta(x, t) = -\delta \cdot \nabla \left[\ln \left(v^\delta(x, t) \right) \right] = -\delta \cdot \frac{\nabla v^\delta(x, t)}{v^\delta(x, t)}. \tag{13}$$

Loosely speaking, (13) shows we can approximate g^k in Algorithm 1 using convolutions with the heat kernel $\Phi_{\delta t}$. A more practical formula for the convolutions in (13) is given in Section 5.

4 Time Evolution

To prevent the sequence $\{x^k\}$ from converging to local minima that are not globally optimal, the HJ equation is evolved in time. The time stepping method we propose is similar in spirit to trust region methods. If the gradient ∇u is small, then time is increased to get out of non-global local minima. If the gradient is relatively large, then time is reduced to utilize the local landscape. In mathematical terms, for $0 < \eta_- < 1 < \eta_+$ and $0 < \theta_1 \leq \theta_2 < 1$, and $\varepsilon > 0$, the time stepping rule is defined by

$$\text{TimeStep}(t, p, q; \tau, T) \triangleq \begin{cases} \min \left(\eta_+ t, T \right) & \text{if } \|p\| \leq \theta_1 \|q\| + \varepsilon \\ t & \text{else if } \|p\| \leq \theta_2 \|q\| + \varepsilon \\ \max \left(\eta_- t, \tau \right) & \text{otherwise.} \end{cases} \tag{14}$$

Algorithms 1 and 2 evolve time using the above time stepping rule with $t = t_k$, $p = g^k$ and $q = g^{k-1}$.

Remark 4.1. *Other updates can be used to generate $\{t_k\}$. We restrict our presentation to that above for simplicity. Future work may investigate convergence improvements with other time step rules.*

5 Gradient Estimation

Expectation Formulation   An essential property of any optimization algorithm is it scales well with dimension n. At first glance, the expression for ∇u^δ in (13) consists of two convolutions, which require evaluating f over all of $\mathbb{R}^n$. This is an intractable task; however, the heat kernel coincides with the probability density of a Gaussian distribution with mean x and standard deviation $\sqrt{\delta t}$. This enables the convolutions to be written as expectations, which can be approximated via sampling (*e.g.* see [20] for an overview of sampling methods). Namely, setting $\mathbb{P}_{x,t} \triangleq \mathcal{N}(x, \delta t)$ yields[3]

$$\nabla u^\delta(x, t) = \frac{1}{t} \cdot \left(x - \frac{\mathbb{E}_{y \sim \mathbb{P}_{x,\delta t}} \left[y \cdot \exp \left(-\delta^{-1} f(y) \right) \right]}{\mathbb{E}_{y \sim \mathbb{P}_{x,\delta t}} \left[\exp \left(-\delta^{-1} f(y) \right) \right]} \right). \tag{15}$$

When t is small, a few samples are needed to adequately estimate the expectation in (15) since the heat kernel is concentrated for small time. When t is large, many samples are needed. However, it is worth noting no grids are required when solving the HJ equation [32, 37, 38].

Exponentially Weighted Moving Average   To reduce the number of samples needed to adequately estimate gradients, especially when n and/or t are large, moving averages of gradients can be used. This allows samples from previous steps to estimate the gradient at the current step. This is similar to ADAM's [25] variance reduction, which estimates first moments via a moving average of gradients.

Limitations   Some limitations of HJ-MAD require careful consideration. First, since HJ-MAD relies on smoothing the objective function, it is expected to not perform as well on functions that are very flat and smooth. Second, while the method is guaranteed to converge to the global minimizer, it does not necessarily outperform some of the state-of-the-art methods. We simply present a method with theoretical guarantees, whereas several current methods are mostly heuristic. Finally, as in most algorithms, the choice of parameters (in this case, time update parameters) is problem-dependent.

---

[3] For completeness, we include a brief derivation of this connection in Appendix B.

6 Related Works

Global Optimization Algorithms Random Search Methods [39] are derivative-free methods that iteratively move to better positions in the search-space. Pure Random Search (PRS) [6] samples points from the domain independently and set the point with lowest function value to be the next iterate. Differential Evolution (DE) [44] uses a population/batch of candidate solutions that are moved around using a simple formula. If the new position of an agent is an improvement then it is accepted and forms part of the population, otherwise the new position is simply discarded. The method is iterative and heuristic. It is similar to PRS and requires many function evaluations. Moreover, there are no convergence guarantees. Basin-hopping (BH) [36, 46] is a two-phase method that iterates by performing random perturbation of coordinates, performing local optimization, and accepting or rejecting new coordinates based on a minimized function value. There are no convergence guarantees. Simulated Annealing [26] is a stochastic, heuristic method that starts at a randomized point x in the parameter space, and then evaluates a neighboring point x', usually chosen at random. In its simplest form, if the value of the objective function is lesser at the new point, the new point is accepted, and the process is repeated. If the value at the new point is greater, the point is chosen with some acceptance probability P; this allows the currently best point considered by the algorithm to zero in on optima in f, but to escape local optima. A method with comparable convergence guarantees to our work is the DIRECT (DIviding RECTangles) [21] method, which partitions a bound-constrained domain into $2n + 1$ hyper-rectangles with an evaluated point at the centre of each. Each hyper-rectangle is scored via a combination of the length of its longest side and the function value at its centre. This scoring favors hyper-rectangles exhibiting both long sides and small function values; the best-scoring hyper-rectangles are further divided.

Moreau Envelope Minimization Our work relates to entropy gradient descent (EGD) [10, 11], which fixes a time T for which to compute the Moreau envelope. There gradients of the Moreau envelope are approximated using a subroutine involving (stochastic) gradient descent. A key difference is HJ-MAD is a zero-order method. Moreover, HJ-MAD does not require a subroutine and instead leverages the Cole-Hopf [18] formula to approximate the gradient via an expectation as in (15). Also, time is adaptive in HJ-MAD whereas it is fixed in EGD. Another work related to ours is the Bend, Mix and Release (BMR), which provides a smooth approximation of the Moreau envelope [41] (albeit without the HJ PDE). Finally, many works also consider minimization of the Moreau envelope, either for weakly convex functions [14, 15], or escaping saddle points of the Moreau envelope [13].

Zero-Order Algorithms HJ-MAD falls under the category of zero-order methods as it does not require gradients of f. In fact, HJ-MAD does not require that f be differentiable. Related methods include the following. Random Gradients [17, 27–29] project gradients onto a random subspace. ZORO [9] assumes sparsity of the gradient and is aimed at high-dimensional problems. ZO-BCD [7] provides a sub-linear query and per-iteration computational complexity. NOMAD is a mesh-adaptive direct search algorithm which are based on progressive-barrier or filter approaches to deal with constraint inequalities and has similar convergence properties as HJ-MAD [31]. Other zero-order methods include derivative-free quasi-Newton methods [4, 30, 33], finite-difference-based methods [42, 43], numerical quadrature-based methods [1, 24], Bayesian methods [30], and comparison methods [8].

7 Experiments

We test HJ-MAD on a set of non-convex benchmark test functions obtained from the Virtual Library of Simulation Experiments [45]. All experiments were run via Google Colaboratory [5]. We start by showing the efficacy of HJ-MAD on the highly nonconvex 2D Griewank function

$$f(x) \triangleq 1 + \sum_{i=1}^{n} \frac{x_i^2}{4000} - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right), \tag{16}$$

which has many widespread local minima. Optimization paths are shown in Figure 2 for HJ-MAD and Gradient Descent (GD). For HJ-MAD, we use 100 samples to estimate the derivative of the Moreau envelope according to (15). As gradient descent is a local optimization algorithm, it converges to a local minimum while HJ-MAD converges to the global minimizer.

(a) Objective Value vs. Iteration   (b) Relative Error vs. Iteration   (c) Optimization Paths
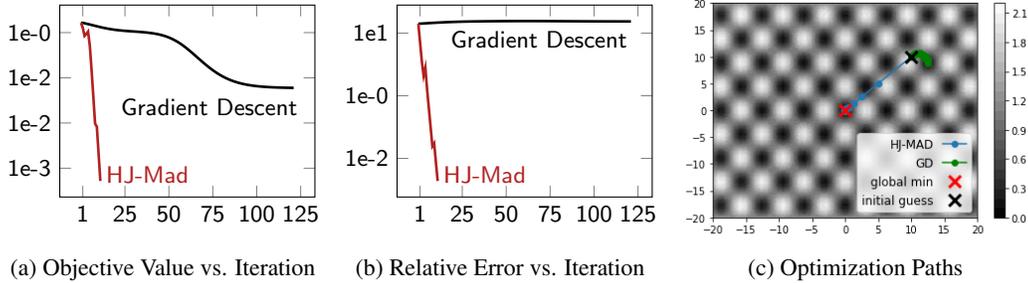
Figure 2: Objective values (left), relative errors (center), and optimization paths (right) for HJ-MAD. HJ-MAD converges to the global minimum while gradient descent converges to a local minimum.

As a more extensive experiment, we compare HJ-MAD with a series of global optimization algorithms mentioned in Section 6. In particular, we compare HJ-MAD with built-in global optimization algorithms from the Python-based package SciPy [22]. These algorithms include PRS [39], DE [44], BH [36, 46], and Dual Annealing [26]. These algorithms are tested on a series of non-convex benchmark test functions obtained from the Virtual Library of Simulation Experiments [45]. Description of these functions can be found in [5]. In Table 1, we list the number of function (and gradient if used) evaluations to get to convergence (*i.e.* within tolerance 5×10^{-2} of the global minimum). Our results show HJ-MAD converges for every test function.

	HJ-MAD	PRS	DE [44]	BH [36]	Annealing [26]
Griewank	167	460K	N	N	451.4K
Drop-Wave	9111	52.5K	1152	N	485.8K
Alpine N.1	635	755.6K	N	N	N
Ackley	498	243.2K	3003	476(116)	3.7M
Levy	5433	N	N	N	N
Rastrigin	500	660.2K	2223	48(12)	590.2K

Table 1: Comparison of global optimization algorithms. Rows show benchmark functions and columns show algorithms. The number in each box gives used function (and gradient in parenthesis) evaluations. "N" means the method did not converge. HJ-MAD results are averaged over 30 trials.

8   Conclusion

We propose a Hamilton-Jacobi-based Moreau Adaptive Descent (HJ-MAD) method for finding global solutions to optimization problems. HJ-MAD is a zero-order algorithm that does not require differentiable objective functions. Our approach is based on two key ideas. First, we use the Moreau envelope with sufficiently large time t such that its minimizers are global minimizers of the objective function. Second, we leverage connections with Hamilton-Jacobi equations to obtain analytic expressions for the Moreau envelope and its gradient. In particular, we use the Cole-Hopf and Hopf-Lax formulas to express the gradient of the Moreau envelope as an expectation. To make the sampling of these expectations more efficient, we include an adaptive time-stepping scheme. Our work also provides a way to estimate proximal operators, in general, using connections to HJ equations. In our experiments, HJ-MAD is the most efficient algorithm and always manages to converge to the global minimizer. Future work may improve the efficiency of HJ-MAD.

Acknowledgements

References

[1] L. B. Almeida. A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In *Artificial neural networks: concept learning*, pages 102–111. 1990.

[2] H. H. Bauschke, P. L. Combettes, et al. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2nd edition, 2017.

[3] A. Beck. *First-Order Methods in Optimization*. SIAM, 2017.

[4] A. S. Berahas, R. H. Byrd, and J. Nocedal. Derivative-free optimization of noisy functions via quasi-newton methods. *SIAM Journal on Optimization*, 29(2):965–993, 2019.

[5] E. Bisong. Google colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pages 59–64. Springer, 2019.

[6] S. H. Brooks. A discussion of random methods for seeking maxima. *Operations research*, 6(2):244–251, 1958.

[7] H. Cai, Y. Lou, D. McKenzie, and W. Yin. A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization. In *International Conference on Machine Learning*, pages 1193–1203. PMLR, 2021.

[8] H. Cai, D. Mckenzie, W. Yin, and Z. Zhang. A one-bit, comparison-based gradient estimator. *Applied and Computational Harmonic Analysis*, 60:242–266, 2022.

[9] H. Cai, D. Mckenzie, W. Yin, and Z. Zhang. Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling. *SIAM Journal on Optimization*, 32(2):687–714, 2022.

[10] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[11] P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):1–30, 2018.

[12] M. G. Crandall and P.-L. Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Mathematics of computation*, 43(167):1–19, 1984.

[13] D. Davis, M. Díaz, and D. Drusvyatskiy. Escaping strict saddle points of the Moreau envelope in nonsmooth optimization. *SIAM Journal on Optimization*, 32(3):1958–1983, 2022.

[14] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

[15] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[16] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018.

[17] Y. M. Ermoliev and R.-B. Wets. *Numerical techniques for stochastic optimization*. Springer-Verlag, 1988.

[18] L. C. Evans. Partial Differential Equations. *Graduate Studies in Mathematics*, 19, 2010.

[19] L. C. Evans. Envelopes and nonconvex Hamilton–Jacobi equations. *Calculus of Variations and Partial Differential Equations*, 50(1):257–282, 2014.

[20] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[21] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of optimization Theory and Applications*, 79(1):157–181, 1993.

[22] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

[23] A. Jourani, L. Thibault, and D. Zagrodny. Differential properties of the Moreau envelope. *Journal of Functional Analysis*, 266(3):1185–1237, 2014.

[24] B. Kim, H. Cai, D. McKenzie, and W. Yin. Curvature-aware derivative-free optimization. *arXiv preprint arXiv:2109.13391*, 2021.

[25] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[26] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[27] D. Kozak, S. Becker, A. Doostan, and L. Tenorio. Stochastic subspace descent. *arXiv preprint arXiv:1904.01145*, 2019.

[28] D. Kozak, S. Becker, A. Doostan, and L. Tenorio. A stochastic subspace approach to gradient-free optimization in high dimensions. *Computational Optimization and Applications*, 79(2):339–368, 2021.

[29] D. Kozak, C. Molinari, L. Rosasco, L. Tenorio, and S. Villa. Zeroth order optimization with orthogonal random directions. *arXiv preprint arXiv:2107.03941*, 2021.

[30] J. Larson, M. Menickelly, and S. M. Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

[31] S. Le Digabel. Algorithm 909: Nomad: Nonlinear optimization with the mads algorithm. *ACM Transactions on Mathematical Software (TOMS)*, 37(4):1–15, 2011.

[32] X. D. Liu, S. Osher, and T. Chan. Weighted Essentially Non-Oscillatory Schemes. *Journal of computational physics*, 115(1):200–212, 1994.

[33] J. Moré and S. Wild. Benchmarking derivative-free optimization algorithms. *SIAM Journal on Optimization*, 20(1):172–191, 2009.

[34] J. J. Moreau. Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 255:238–240, 1962.

[35] J. J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[36] B. Olson, I. Hashmi, K. Molloy, and A. Shehu. Basin hopping as a general and versatile optimization framework for the characterization of biological macromolecules. *Advances in Artificial Intelligence (16877470)*, 2012.

[37] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of computational physics*, 79(1):12–49, 1988.

[38] S. Osher and C.-W. Shu. High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations. *SIAM Journal on numerical analysis*, 28(4):907–922, 1991.

[39] L. Rastrigin. The convergence of the random search method in the extremal control of a many parameter system. *Automaton & Remote Control*, 24:1337–1342, 1963.

[40] R. T. Rockafellar. *Convex Analysis*, volume 18. Princeton University Press, 1970.

[41] K. Scaman, L. Dos Santos, M. Barlier, and I. Colin. A simple and efficient smoothing method for faster optimization and local exploration. *Advances in Neural Information Processing Systems*, 33:6503–6513, 2020.

[42] H.-J. M. Shi, Y. Xie, M. Q. Xuan, and J. Nocedal. Adaptive finite-difference interval estimation for noisy derivative-free optimization. *arXiv preprint arXiv:2110.06380*, 2021.

[43] H.-J. M. Shi, M. Q. Xuan, F. Oztoprak, and J. Nocedal. On the numerical performance of derivative-free optimization methods based on finite-difference approximations. *arXiv preprint arXiv:2102.09762*, 2021.

[44] R. Storn and K. Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[45] S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets. Retrieved June 23, 2021, from http://www.sfu.ca/~ssurjano.

[46] D. J. Wales and J. P. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997.

A Proofs

Below a sequence of lemmas is provided to obtain the main result. The first lemma is an elementary result, a minor tweak on known results (*e.g.* the early works [34, 35] about proximals being nonempty).

Lemma A.1. *Given time $t \in (0, \infty)$, if Assumptions 2.1 and 2.2 hold for $f \colon \mathbb{R}^n \to \mathbb{R}$, then $\mathrm{prox}_{tf}(x)$ is nonempty for all $x \in \mathbb{R}^n$.*

Proof. Fix $x \in \mathbb{R}^n$. Consider the set

$$\mathcal{T} \triangleq \left\{ z : f(z) + \frac{1}{2t}\|z - x\|^2 \leq f(x) + 1 \right\}. \tag{17}$$

By Assumption 2.2i and Assumption 2.1, the extreme value theorem asserts f attains its minimum value on $\mathcal{S}_\gamma$, and so there is a global minimizer $x^\star$ of f. Consequently, the constraint in $\mathcal{T}$ can be rewritten to obtain

$$\|z - x\|^2 \leq 2t(f(x) - f(z) + 1) \leq 2t(f(x) - f(x^\star) + 1), \tag{18}$$

and so $\mathcal{T} \subseteq B(x, 2t[f(x) - f(x^\star) + 1])$, *i.e.* $\mathcal{T}$ is bounded. Moreover, since $x \in \mathcal{T}$, $\mathcal{T}$ is nonempty, and $\mathcal{T}$ is closed because f is continuous. Due to the fact the minimal value of f is a lower bound, *i.e.*

$$f(z) + \frac{1}{2t}\|z - x\|^2 \geq f(x^\star) \quad \text{for all } z \in \mathbb{R}^n, \tag{19}$$

the set

$$\left\{ f(z) + \frac{1}{2t}\|z - x\|^2 : z \in \mathbb{R}^n \right\} \tag{20}$$

has an infimum. Since $x \in \mathcal{T}$, this infimum is less than or equal to $f(x)$. Thus, there exists a sequence $\{z^k\} \subset \mathbb{R}^n$ such that

$$\lim_{k \to \infty} f(z^k) + \frac{1}{2t}\|z^k - x\|^2 = \inf_{z \in \mathbb{R}^n} f(z) + \frac{1}{2t}\|z - x\|^2. \tag{21}$$

By definition of this limit and the fact the infimum does not exceed $f(x)$, there is $N \in \mathbb{N}$ such that

$$f(z^k) + \frac{1}{2t}\|z^k - x\|^2 \leq \left(\inf_{z \in \mathbb{R}^n} f(z) + \frac{1}{2t}\|z - x\|^2 \right) + 1 \leq f(x) + 1, \quad \text{for all } k \geq N, \tag{22}$$

and so $z^k \in \mathcal{T}$ for $k \geq N$. This implies the sequence $\{z^k\}$ is bounded, whereby the Bolzano–Weierstrass theorem may be applied to deduce existence of a convergent subsequence $\{z^{n_k}\} \subseteq \{z^k\}$ with limit $\hat{x} \in \mathcal{T}$ (*n.b.* $\mathcal{T}$ is closed). Then observe

$$f(\hat{x}) + \frac{1}{2t}\|\hat{x} - x\|^2 = \lim_{k \to \infty} f(z^{n_k}) + \frac{1}{2t}\|z^{n_k} - x\|^2 = \inf_{z \in \mathbb{R}^n} f(z) + \frac{1}{2t}\|z - x\|^2, \tag{23}$$

which implies $\hat{x} \in \mathrm{prox}_{tf}(x)$, and the proof is complete. $\qquad\square$

Below is a nearly trivial result that is also well-known for Moreau envelopes.

Lemma A.2. *Given a time $t \in (0, \infty)$, if Assumptions 2.1 and 2.2 hold for $f \colon \mathbb{R}^n \to \mathbb{R}$, then $0 < t \leq T$ implies*

$$u(x, T) \leq u(x, t) \quad \text{for all } x \in \mathbb{R}^n. \tag{24}$$

Proof. By Lemma A.1, $\hat{x} \in \text{prox}_{tf}(x)$ exists, and we have

$$u(x, T) \leq f(\hat{x}) + \frac{1}{2T}\|\hat{x} - x\|^2 \leq f(\hat{x}) + \frac{1}{2t}\|\hat{x} - x\|^2 = u(x, t), \quad \text{for all } x \in \mathbb{R}^n. \tag{25}$$

$\square$

Below is a gradient $\nabla u(x, t)$ formula *without* assuming f is convex. This is a modification of Proposition 3.1 in [23]. This is a minor extension of a well-known result in convex analysis originating from Moreau [35] (*e.g.* see Proposition 12.30 in [2], Theorem 6.60 in [3], and Thereom 31.5 in [40].)

Lemma A.3. *Given a time $t \in (0, \infty)$, if Assumptions 2.1 and 2.2 hold for $f \colon \mathbb{R}^n \to \mathbb{R}$ and if the envelope $u(\cdot, t)$ of f is differentiable at $x \in \mathbb{R}^n$, then $\text{prox}_{tf}(x)$ is a singleton set and*

$$\nabla u(x, t) = \frac{x - \hat{x}}{t}, \quad \text{where} \quad \hat{x} \triangleq \text{prox}_{tf}(x) = \underset{z \in \mathbb{R}^n}{\text{argmin}}\, f(z) + \frac{1}{2t}\|z - x\|^2. \tag{26}$$

Proof. By Lemma A.1, there exists $\hat{x} \in \text{prox}_{tf}(x)$. Fix any $h \in \mathbb{R}^n$. By definition of the gradient,

$$\langle \nabla u(x, t), h \rangle = \lim_{\lambda \to 0} \frac{u(x, t) - u(x - \lambda h, t)}{\lambda} \tag{27a}$$

$$= \lim_{\lambda \to 0} \frac{f(\hat{x}) + \frac{1}{2t}\|\hat{x} - x\|^2 - u(x - \lambda h, t)}{\lambda} \tag{27b}$$

$$\geq \lim_{\lambda \to 0} \frac{f(\hat{x}) + \frac{1}{2t}\|\hat{x} - x\|^2 - f(\hat{x}) - \frac{1}{2t}\|\hat{x} - (x - \lambda h)\|^2}{\lambda} \tag{27c}$$

$$= \lim_{\lambda \to 0} \frac{\frac{1}{t}\langle \hat{x} - x, -\lambda h \rangle - \frac{1}{2t}\|\lambda h\|^2}{\lambda} \tag{27d}$$

$$= \left\langle \frac{x - \hat{x}}{t}, h \right\rangle, \tag{27e}$$

where the equality (27b) holds because $\hat{x} \in \text{prox}_{tf}(x)$ and the subsequent inequality (27c) follows from the fact $u(x - \lambda h, t)$ is the infimum of $f(z) + \frac{1}{2t}\|z - (x - \lambda h)\|^2$ among all $z \in \mathbb{R}^n$. Since (27) holds for arbitrarily chosen h, this limit inequality holds over $\mathbb{R}^n$. Consequently,

$$\left\langle \nabla u(x, t) - \frac{x - \hat{x}}{t}, z \right\rangle \geq 0, \quad \text{for all } z \in \mathbb{R}^n. \tag{28}$$

By way of contradiction, suppose the gradient formula in (26) does not hold, *i.e.*

$$\nabla u(x, t) - \frac{x - \hat{x}}{t} \neq 0. \tag{29}$$

In such a case, taking

$$z = \frac{x - \hat{x}}{t} - \nabla u(x, t) \tag{30}$$

yields

$$0 \leq \left\langle \nabla u(x, t) - \frac{x - \hat{x}}{t}, z \right\rangle = -\left\| \nabla u(x, t) - \frac{x - \hat{x}}{t} \right\|^2 < 0, \tag{31}$$

which implies $0 < 0$, a contradiction. Thus, (29) must be false, from which the result (26) follows.

All that remains is to verify $\hat{x}$ is the unique element of $\mathrm{prox}_{tf}(x)$. To this end, fix any $\hat{y} \in \mathrm{prox}_{tf}(x)$ and observe, repeating the same argument as for the gradient formula above,

$$\frac{x - \hat{y}}{t} = \nabla u(x, t) = \frac{x - \hat{x}}{t} \quad \Longleftrightarrow \quad \hat{x} = \hat{y}. \tag{32}$$

That is, $\hat{y} \in \mathrm{prox}_{tf}(x)$ if and only if $\hat{y} = \hat{x}$, completing the proof. $\square$

The following lemma is again a generalization of well-known results in convex analysis; here again the proof adapts arguments from [23].

Lemma A.4. *Given a time* $t \in (0, \infty)$, *if Assumptions 2.1 and 2.2 holds for* $f \colon \mathbb{R}^n \to \mathbb{R}$ *and if* $x \in \mathbb{R}^n$ *is a local minimizer of the envelope* $u(\cdot, t)$ *of* f, *then* x *is a local minimizer of* f, $u(x, t) = f(x)$, $x = \mathrm{prox}_{tf}(x)$, $u(\cdot, t)$ *is differentiable at* x, *and*

$$\nabla u(x, t) = 0. \tag{33}$$

Proof. Suppose x is a local minimizer of $u(\cdot, t)$. By Lemma A.1, there exists $\hat{x} \in \mathrm{prox}_{tf}(x)$. We first show $x = \hat{x}$ (Step 1), which implies $u(x, t) = f(x)$. This is used to verify $\nabla u(x, t) = 0$ (Step 2), and we conclude by showing x is a local minimizer of f (Step 3).

Step 1. Fix any $h \in \mathbb{R}^n$. Since x is a local minimizer of u, there is $\overline{\lambda} > 0$ such that, for $\lambda \in (-\overline{\lambda}, \overline{\lambda})$,

$$u(x - \lambda h, t) \leq f(\hat{x}) + \frac{1}{2t}\|\hat{x} - (x - \lambda h)\|^2 \tag{34a}$$

$$= f(\hat{x}) + \frac{1}{2t}\|\hat{x} - x\|^2 + \left\langle \frac{\hat{x} - x}{t}, \lambda h \right\rangle + \frac{\lambda^2}{2t}\|h\|^2 \tag{34b}$$

$$= u(x, t) + \left\langle \frac{\hat{x} - x}{t}, \lambda h \right\rangle + \frac{\lambda^2}{2t}\|h\|^2. \tag{34c}$$

Upon rearrangement, we deduce

$$\left\langle \frac{x - \hat{x}}{t}, \lambda h \right\rangle - \frac{\lambda^2}{2t}\|h\|^2 \leq u(x, t) - u(x - \lambda h, t) \leq 0, \quad \text{for all } \lambda \in (-\overline{\lambda}, \overline{\lambda}). \tag{35}$$

This implies

$$0 \geq \lim_{\lambda \to 0^+} \frac{u(x, t) - u(x - \lambda h, t)}{\lambda} \geq \lim_{\lambda \to 0^+} \left\langle \frac{x - \hat{x}}{t}, h \right\rangle - \frac{\lambda}{2t}\|h\|^2 = \left\langle \frac{x - \hat{x}}{t}, h \right\rangle. \tag{36}$$

By the arbitrariness of h, we may choose

$$h = \frac{\hat{x} - x}{t} \tag{37}$$

and apply (36) to find

$$\left\| \frac{x - \hat{x}}{t} \right\|^2 \leq 0 \quad \Longrightarrow \quad x = \hat{x}. \tag{38}$$

Step 2. Plugging $x = \hat{x}$ back into (36) furthermore reveals, by the squeeze lemma,

$$\langle \nabla u(x, t), h \rangle = \lim_{\lambda \to 0} \frac{u(x, t) - u(x - \lambda h, t)}{\lambda} = 0, \quad \text{for all } h \in \mathbb{R}^n. \tag{39}$$

Since h was arbitrarily chosen, it follows that $\nabla u(x, t) = 0$.

Step 3. By way of contradiction, suppose x is not a local minimizer of f. This would imply existence of nonzero $h \in \mathbb{R}^n$ and $\hat{\lambda} \in (0, \infty)$ such that

$$f(x + \lambda h) < f(x) \quad \text{for all } \lambda \in (0, \hat{\lambda}). \tag{40}$$

However, our above results together with the truth of (40) would imply

$$u(x + \lambda h, t) \leq f(x + \lambda h) + \frac{1}{2t}\|(x + \lambda h) - (x + \lambda h)\|^2 \tag{41a}$$

$$= f(x + \lambda h, t) \tag{41b}$$

$$< f(x) \tag{41c}$$

$$= u(x, t), \quad \text{for all } \lambda \in (0, \hat{\lambda}), \tag{41d}$$

contradicting the fact x is a local minimizer of $u(\cdot, t)$. Thus, x is a local minimizer of f. $\square$

The lemma below is widely known in the case where f is convex. For completeness, we include its adaptation to our setting with a more general subdifferential definition and unique assumptions.

Lemma A.5. *Given a time $t \in (0, \infty)$, if Assumptions 2.1 and 2.2 hold for $f \colon \mathbb{R}^n \to \mathbb{R}$ and if there is $x \in \mathbb{R}^n$ such that $f(x) = u(x, t)$, then $0 \in \partial f(x)$.*

Proof. By Lemma A.1, there is $\hat{x} \in \text{prox}_{tf}(x)$, and so

$$f(\hat{x}) + \frac{1}{2t}\|\hat{x} - x\|^2 \le f(z) + \frac{1}{2t}\|z - x\|^2, \quad \text{for all } z \in \mathbb{R}^n. \tag{42}$$

By definition of u and the given hypothesis,

$$f(x) = u(x, t) = f(\hat{x}) + \frac{1}{2t}\|\hat{x} - x\|^2. \tag{43}$$

Combining (42) and (43) reveals

$$f(x) + \langle 0, z - x \rangle - \frac{1}{2t}\|z - x\|^2 \le f(z), \quad \text{for all } z \in \mathbb{R}^n. \tag{44}$$

Since $\|z - x\|^2 = o(\|z - x\|)$ as $z \to x$, (44) shows, by Definition 2.1, $0 \in \partial f(x)$. $\qquad\square$

The next two lemmas are additional auxiliary results we introduce to utilize our unique assumptions.

Lemma A.6. *Given a time $t \in (0, \infty)$, if Assumptions 2.1 and 2.2 hold for $f \colon \mathbb{R}^n \to \mathbb{R}$ and if $x \in \mathbb{R}^n$ is a local minimizer of the envelope $u(\cdot, t)$ of f and $t \ge \|x - x^\star\|^2 / 2\gamma$ for some global minimizer $x^\star$ of f, then x is a global minimizer of f.*

Proof. Due to Assumption 2.2i and Assumption 2.1, the extreme value theorem asserts f attains its minimum value on $\mathcal{S}_\gamma$, *i.e.* there is a global minimizer $x^\star$ of f. By Lemma A.4, $f(x) = u(x, t)$, and

$$f(x) = u(x, t) \le f(x^\star) + \frac{1}{2t}\|x^\star - x\|^2 \le f(x^\star) + \gamma, \tag{45}$$

i.e. $x \in \mathcal{S}_\gamma$. By Lemma A.5, $0 \in \partial f(x)$. These facts together with Assumption 2.2ii imply x is a global minimizer of f. $\qquad\square$

Lemma A.7. *Given a time $t \in (0, \infty)$, if Assumptions 2.1, 2.2 and 2.3 hold for $f \colon \mathbb{R}^n \to \mathbb{R}$ and if the sequence $\{u(x^k, t_k)\}$ is monotonically decreasing, then the sequence $\{x^k\}$ is bounded.*

Proof. By Assumption 2.3ii, there is global minimizer $x^\star$ of f. Define the set $\mathcal{Q} \triangleq \{z : u(z, T) \le f(x^\star) + \gamma\}$. We first show $\mathcal{Q}$ is bounded (Step 1) and then that $x^k \in \mathcal{Q}$ for all $k \in \mathbb{N}$ (Step 2).

Step 1. Since $x^\star$ is a global minimizer, f is bounded from below. By way of contradiction, suppose $\mathcal{Q}$ is unbounded. This implies there is a sequence $\{z^k\} \subseteq \mathcal{Q}$ such that

$$\lim_{k \to \infty} \|z^k\| = +\infty. \tag{46}$$

By Lemma A.1, there is a sequence $\{\hat{z}^k\}$ such that $\hat{z}^k \in \text{prox}_{Tf}(z^k)$, for all $k \in \mathbb{N}$, and

$$f(\hat{z}^k) \le f(\hat{z}^k) + \frac{1}{2T}\|\hat{z}^k - z^k\|^2 = u(z^k, T) \le f(x^\star) + \gamma \implies \hat{z}^k \in \mathcal{S}_\gamma, \tag{47}$$

where the final inequality holds since $z^k \in \mathcal{Q}$. This implies $\{\hat{z}^k\}$ is bounded by $B > 0$ and

$$\|\hat{z}^k - z^k\|^2 \le 2T\left(f(x^\star) - f(\hat{z}^k) + \gamma\right) \le 2T\gamma, \quad \text{for all } k \in \mathbb{N}. \tag{48}$$

Thus,

$$\infty = \lim_{k \to \infty} \|z^k\| \le \lim_{k \to \infty} \|\hat{z}^k\| + \|\hat{z}^k - z^k\| \le \lim_{k \to \infty} B + \sqrt{2T\gamma}, \tag{49}$$

a contradiction.

Step 2. By the monotonicity of $\{u(x^k, t_k)\}$, Assumption 2.3ii, and Lemma A.2,

$$u(x^k, T) \le u(x^k, t_k) \le u(x^1, t_1) \le f(x^\star) + \frac{1}{2t_1}\|x^\star - x^1\|^2 \le f(x^\star) + \gamma, \quad \text{for all } k \in \mathbb{N}, \tag{50}$$

which implies $\{x^k\} \subseteq \mathcal{Q}$. $\qquad\square$

The following lemma is a novel result using the assumptions of our setting; it draws inspiration from inequalities in the analysis in Theorem 2.1 of [14].

Lemma A.8. *Given a time $t \in (0, \infty)$, if Assumptions 2.1, 2.2 and 2.3 hold for $f \colon \mathbb{R}^n \to \mathbb{R}$ and if $\{x^k\}$ is generated by the iteration in Lines 3 to 6 of Algorithm 1, then the sequence $\{u(x^k, t_k)\}$ monotonically decreases and converges and there is $\mu > 0$ such that*

$$u(x^{k+1}, t_{k+1}) \leq u(x^k, t_k) - \frac{\mu}{2}\|x^k - \hat{x}^k\|^2, \quad \text{for all } k \in \mathbb{N}, \tag{51}$$

where $\hat{x}^k \in \mathrm{prox}_{t_k f}(x^k)$ for all $k \in \mathbb{N}$.

Proof. By Lemma A.1, there is $\hat{x}^k \in \mathrm{prox}_{t_k f}(x^k)$ for all $k \in \mathbb{N}$. Additionally, for all $k \in \mathbb{N}$,

$$u(x^{k+1}, t_{k+1}) \leq f(\hat{x}^k) + \frac{1}{2t_{k+1}}\|\hat{x}^k - x^{k+1}\|^2 \tag{52a}$$

$$= f(\hat{x}^k) + \frac{1}{2t_{k+1}}\|\hat{x}^k - x^k + \alpha t_k g^k\|^2 \tag{52b}$$

$$= f(\hat{x}^k) + \frac{(1-\alpha)^2}{2t_{k+1}}\|x^k - \hat{x}^k\|^2 \tag{52c}$$

$$= u(x^k, t_k) + \frac{1}{2}\underbrace{\left(\frac{(1-\alpha)^2}{t_{k+1}} - \frac{1}{t_k}\right)}_{\triangleq \xi_k}\|x^k - \hat{x}^k\|^2, \tag{52d}$$

where ξ_k is defined to be the underbraced quantity. The first inequality above follows from the definition of u and the second equality holds by definition of the update formula for $\{x^k\}$. We next show ξ_k is bounded from above by a negative constant. Define the sequence $\{\theta_k\} \subset (0, \infty)$ by

$$\theta_k \triangleq \frac{t_{k+1}}{t_k}, \quad \text{for all } k \in \mathbb{N}. \tag{53}$$

Note η_- is a lower bound for $\{\theta_k\}$ and, by the choice of step size in Assumption 2.3, $(1-\alpha_k)^2 < \eta_-$. Thus, there is $\mu \in (0, 1/T)$ such that $(1-\alpha)^2 \leq (1 - \mu T)\eta_-$ and

$$\theta_k \geq \eta_- \geq \frac{(1-\alpha)^2}{1 - \mu T} \geq \frac{(1-\alpha)^2}{1 - \mu t_k}, \quad \text{for all } k \in \mathbb{N}, \tag{54}$$

which implies

$$1 - \mu t_k \geq \frac{(1-\alpha)^2}{\theta_k} \implies -\mu \geq \frac{1}{t_k}\left(\frac{(1-\alpha)^2}{\theta_k} - 1\right) = \frac{(1-\alpha)^2}{t_{k+1}} - \frac{1}{t_k}, \quad \text{for all } k \in \mathbb{N}. \tag{55}$$

Together (52) and (55) imply

$$u(x^{k+1}, t_{k+1}) \leq u(x^k, t_k) - \frac{\mu}{2}\|x^k - \hat{x}^k\|^2, \quad \text{for all } k \in \mathbb{N}, \tag{56}$$

i.e. $\{u(x^k, t_k)\}$ is monotonically decreasing. Then (51) follows from induction on (56). Additionally, Assumption 2.3 implies there is a global minimizer $x^\star$ of f. Whence

$$-\infty < f(x^\star) \leq \inf_x f(x) + \frac{1}{2t_k}\|x - x^k\|^2 = u(x^k, t_k), \quad \text{for all } k \in \mathbb{N}. \tag{57}$$

By monotone convergence theorem, the sequence $\{u(x^k, t_k)\}$ converges. $\qquad \square$

We conclude this section below with our main contribution, a convergence theorem.

Theorem 2.1. (GLOBAL MINIMIZATION). If Assumptions 2.1, 2.2, and 2.3 hold, then the iteration in Lines 4 to 7 of Algorithm 1 yields convergence to optimal objective values, *i.e.*

$$\lim_{k\to\infty} f(x^k) = \min_{x\in\mathbb{R}^n} f(x). \tag{58}$$

Additionally, a subsequence $\{x^{n_k}\} \subseteq \{x^k\}$ converges to a global minimizer of f.

Proof. We first show there is a subsequence $\{x^{n_k}\} \subseteq \{x^k\}$ with limit x^∞ (Step 1). Then we show $\|x^k - \hat{x}^k\| \to 0$, where $\hat{x}^k \in \text{prox}_{t_k f}(x^k)$ (Step 2). This is used to verify $t_k \to T$ in finitely many steps (Step 3). These facts are together used to show $u(x^\infty, T) = f(x^\infty)$ (Step 4). This, in turn, is used to verify x^∞ is a global minimizer of f (Step 5) and $u(x^k, T) \to f(x^\infty)$ (Step 6). We conclude by showing convergence of the function values $\{f(x^k)\}$ to $f(x^\infty)$ (Step 7).

Step 1. By Lemma A.8, $\{u(x^k, t_k)\}$ converges monotonically (*i.e.* is decreasing) and there is $\mu > 0$ and a sequence $\{\hat{x}^k\}$ such that (51) holds and $\hat{x}^k \in \text{prox}_{t_k f}(x^k)$ for all $k \in \mathbb{N}$. The monotonicity of $\{u(x^k, t_k)\}$ with Lemma A.7 implies $\{x^k\}$ is bounded. Thus, there is a convergent subsequence $\{x^{n_k}\} \subseteq \{x^k\}$ with limit x^∞.

Step 2. Due to Assumption 2.3ii, there is a global minimizer $x^\star$ of f, which implies f is bounded from below. If

$$\sum_{n=1}^{\infty} \|x^n - \hat{x}^n\|^2 = +\infty, \tag{59}$$

then (51) implies

$$f(x^\star) = \min_{x\in\mathbb{R}^n} f(x) \leq \lim_{k\to\infty} u(x^k, t_k) \leq \lim_{k\to\infty} u(x^1, t_1) - \frac{\mu}{2}\sum_{n=1}^{\infty} \|x^n - \hat{x}^n\|^2 = -\infty, \tag{60}$$

contradicting the fact f is bounded from below. Thus, the series in (59) is finite. Because the sequence $\{\|x^n - \hat{x}^n\|\}$ is also nonnegative, it further follows that

$$\lim_{k\to\infty} \|x^k - \hat{x}^k\| = 0. \tag{61}$$

Step 3. Note $\{T/t_k\} \subset [1, T/\tau]$ by the choice of time step rule in (14) and, by (61),

$$0 = \lim_{k\to\infty} \frac{T}{t_k}\|x^k - \hat{x}^k\| = T\lim_{k\to\infty} \|g^k\| \implies \lim_{k\to\infty} \|g^k\| = 0, \tag{62}$$

which implies

$$0 \leq \lim_{k\to\infty} \frac{\|g^k\|}{\theta_1\|g^{k-1}\| + \varepsilon} \leq \lim_{k\to\infty} \frac{\|g^k\|}{\varepsilon} = 0. \tag{63}$$

Consequently, by (14), there exists $N_1 \in \mathbb{N}$ such that

$$t_{k+1} = \text{TimeStep}(t_k, g^k, g^{k-1}) = \min(T, \eta_+ t_k), \quad \text{for all } k \geq N_1, \tag{64}$$

and so there exists $N_2 \geq N_1$ such that

$$t_k = T, \quad \text{for all } k \geq N_2. \tag{65}$$

Thus, $t_k \to T$ in finitely many steps.

Step 4. By (61) and the fact $x^{n_k} \to x^\infty$,

$$0 \leq \lim_{k\to\infty} \|\hat{x}^{n_k} - x^\infty\| \leq \lim_{k\to\infty} \|\hat{x}^{n_k} - x^{n_k}\| + \|x^{n_k} - x^\infty\| = 0 + 0, \tag{66}$$

and so the squeeze lemma asserts $\hat{x}^{n_k} \to x^\infty$. Fix any $\beta > 0$. Then the convergence of $\{\hat{x}^{n_k}\}$ implies there is $N_3 \in \mathbb{N}$ such that

$$|f(x^\infty) - f(\hat{x}^{n_k})| \leq \beta, \quad \text{for all } k \geq N_3. \tag{67}$$

By Lemma A.1, there is $\hat{x}^\infty \in \text{prox}_{tf}(x^\infty)$. So, the convergence of $\{x^{n_k}\}$ and continuity of scalar products also implies there is $N_4 \in \mathbb{N}$ such that

$$\left| \|x^{n_k}\|^2 - \|x^\infty\|^2 - 2\langle \hat{x}^\infty, x^{n_k} - x^\infty \rangle \right| \leq 2T\beta, \quad \text{for all } k \geq N_4. \tag{68}$$

14

Together (65), (67), (68) and the definition of u imply, for all $k \geq \max(N_2, N_3, N_4)$,

$$f(x^\infty) \leq f(\hat{x}^{n_k}) + \beta \tag{69a}$$

$$\leq u(x^{n_k}, t_{n_k}) + \beta \tag{69b}$$

$$= u(x^{n_k}, T) + \beta \tag{69c}$$

$$\leq f(\hat{x}^\infty) + \frac{1}{2T}\|\hat{x}^\infty - x^{n_k}\|^2 + \beta \tag{69d}$$

$$= u(x^\infty, T) + \frac{1}{2T}\left[\|x^{n_k}\|^2 - \|x^\infty\|^2 - 2\langle \hat{x}^\infty, x^{n_k} - x^\infty \rangle\right] + \beta \tag{69e}$$

$$\leq u(x^\infty, T) + 2\beta. \tag{69f}$$

Hence

$$u(x^\infty, T) \leq f(x^\infty) \leq u(x^\infty, T) + 2\beta. \tag{70}$$

Since this holds for arbitrary $\beta > 0$, we may let $\beta \to 0^+$ to deduce $f(x^\infty) = u(x^\infty, T)$.

Step 5. Using the convergence of $\{\hat{x}^{n_k}\}$ and monotonicity of $\{u(x^k, t_k)\}$,

$$f(x^\infty) = \lim_{k\to\infty} f(\hat{x}^{n_k}) \leq \lim_{k\to\infty} u(x^{n_k}, t_k) \leq u(x^1, t_1). \tag{71}$$

With Assumption 2.3ii, (71) reveals

$$f(x^\infty) \leq u(x^1, t_1) \leq f(x^\star) + \frac{1}{2t_1}\|x^\star - x^1\|^2 \leq f(x^\star) + \gamma. \tag{72}$$

Additionally, by Lemma A.5 and Step 4, $0 \in \partial f(x^\infty)$. These last two results together with Assumption 2.2ii imply x^∞ is a global minimizer of f.

Step 6. Next we show $u(x^k, T) \to f(x^\infty)$. Since

$$f(x^\infty) = \lim_{k\to\infty} f(\hat{x}^{n_k}) \leq \lim_{k\to\infty} u(x^{n_k}, t_{n_k}) \leq \lim_{k\to\infty} f(x^{n_k}) = f(x^\infty), \tag{73}$$

$u(x^{n_k}, t_k) \to f(x^\infty)$. As $t_k \to T$ in finitely many steps, $u(x^{n_k}, T) \to f(x^\infty)$. Because $\{u(x^{n_k}, T)\} \subseteq \{u(x^k, T)\}$ and $\{u(x^k, T)\}$ converges,[4] the unique limit of the entire sequence must coincide with the limits of subsequences, *i.e.* $u(x^k, T) \to f(x^\infty)$.

Step 7. Let $\varepsilon > 0$ be given. It suffices to show there is $N \in \mathbb{N}$ such that

$$f(x^k) \leq f(x^\infty) + \varepsilon, \quad \text{for all } k \geq N. \tag{74}$$

Since $u(x^k, T) \to f(x^\infty)$, there is $N_5 \in \mathbb{N}$ such that

$$f(\hat{x}^k) \leq u(x^k, T) \leq f(x^\infty) + \min\left(\gamma, \frac{\varepsilon}{2}\right), \quad \text{for all } k \geq N_5. \tag{75}$$

By the continuity of f (*i.e.* Assumption 2.1) and Assumption 2.2i, f is uniformly continuous[5] on the set $\mathcal{C}_\gamma \triangleq \{z : d(z, \mathcal{S}_\gamma) \leq 1\}$. Thus, there is $\delta > 0$ such that, for all $x, y \in \mathcal{C}_\gamma$,

$$\|x - y\| \leq \delta \implies |f(x) - f(y)| \leq \frac{\varepsilon}{2}. \tag{76}$$

Since $\|x^k - \hat{x}^k\| \to 0$, there is $N_6 \in \mathbb{N}$ such that

$$\|x^k - \hat{x}^k\| \leq \min(\delta, 1), \quad \text{for all } k \geq N_6. \tag{77}$$

So, (75) implies $\hat{x}^k \in \mathcal{S}_\gamma \subset \mathcal{C}_\gamma$ for all $k \geq N_5$, and (77) implies $x^{n_k} \in \mathcal{C}_\gamma$ for all $k \geq \max(N_5, N_6)$. Additionally, (75), (76), and (77) together yield

$$f(x^k) \leq f(\hat{x}^k) + \frac{\varepsilon}{2} \tag{78a}$$

$$\leq f(\hat{x}^k) + \frac{1}{2T}\|\hat{x}^k - x^k\|^2 + \frac{\varepsilon}{2} \tag{78b}$$

$$= u(x^k, T) + \frac{\varepsilon}{2} \tag{78c}$$

$$\leq f(x^\infty) + \varepsilon, \quad \text{for all } k \geq \max(N_5, N_6), \tag{78d}$$

which verifies (74), taking $N = \max(N_5, N_6)$. □

---

[4]This holds since $\{u(x^k, t_k)\}$ converges and $u(x^k, t_k) = u(x^k, T)$ for $k \geq N_2$.

[5]The set $\mathcal{S}_\gamma$ is compact and, given the continuity of $d(z, \mathcal{S}_\gamma) \triangleq \inf\{\|z - x\| : x \in \mathcal{S}_\gamma\}$ and boundedness of $\mathcal{S}_\gamma$, the set $\mathcal{C}_\gamma$ is also compact.

B Derivation of Gradient Formula as Expectation

Recall from (13) that the gradient of the viscous HJ solution is

$$\nabla u^{\delta}(x,t) = -\delta \cdot \nabla \left[\ln \left(v^{\delta}(x,t) \right) \right] = -\delta \cdot \frac{\nabla v^{\delta}(x,t)}{v^{\delta}(x,t)}, \tag{79}$$

where the heat equation solution v^{δ} can be rewritten in the form

$$v^{\delta}(x,t) = \left(\Phi_{\delta t} * \exp(-f/\delta) \right)(x) \tag{80}$$

$$= (2\pi\delta t)^{-n/2} \int_{\mathbb{R}^n} \exp\left(\frac{-f(y)}{\delta} \right) \exp\left(\frac{-(x-y)^2}{2\delta t} \right) dy \tag{81}$$

$$= \mathbb{E}_{y \sim \mathbb{P}_{x,\delta t}} \left[\exp\left(-\frac{f(y)}{\delta} \right) \right]. \tag{82}$$

Here, we have re-written the integral as an expectation, noting (81) contains the heat kernel, which could be re-written as a Gaussian density with mean x and standard deviation $\sqrt{\delta t}$. Differentiating v^{δ} with respect to x, we obtain

$$\nabla v^{\delta}(x,t) = (2\pi\delta t)^{-n/2} \int_{\mathbb{R}^n} \frac{x-y}{\delta t} \exp\left(\frac{-(x-y)^2}{2\delta t} \right) \exp\left(\frac{-f(y)}{\delta} \right) dy \tag{83}$$

$$= -\frac{1}{\delta t} \cdot \mathbb{E}_{y \sim \mathbb{P}_{x,\delta t}} \left[(x-y) \exp\left(-\delta^{-1} f(y) \right) \right]. \tag{84}$$

Plugging these definitions of v^{δ} and ∇v^{δ} in (13), we obtain the desired formula in (15).

C Convergence of (12) to Moreau Envelope

Prior work has already established the uniform convergence $u^{\delta} \to u$ as $\delta \to 0^+$. Below we expand the convolution definition and rewrite u^{δ} using an L^p norm. Fixing $x \in \mathbb{R}^n$ and $t > 0$ and defining

$$q_t(y) \triangleq f(y) + \frac{1}{2t} \|y - x\|^2, \tag{85}$$

note

$$u^{\delta}(x,t) = -\delta \ln \left((2\pi\delta t)^{-n/2} \int_{\mathbb{R}^n} \exp\left(\frac{-f(y)}{\delta} \right) \exp\left(\frac{-\|x-y\|^2}{2\delta t} \right) dy \right) \tag{86a}$$

$$= \frac{\delta n}{2} \ln(2\pi\delta t) - \ln \left(\left[\int_{\mathbb{R}^n} \exp\left(-\frac{2tf(y) + \|x-y\|^2}{2\delta t} \right) dy \right]^{\delta} \right) \tag{86b}$$

$$= \frac{\delta n}{2} \ln(2\pi\delta t) - \ln \left(\left[\int_{\mathbb{R}^n} \exp\left(- \left[f(y) + \frac{1}{2t}\|x-y\|^2 \right] \right)^{1/\delta} dy \right]^{\delta} \right) \tag{86c}$$

$$= \frac{\delta n}{2} \ln(2\pi\delta t) - \ln \left(\left[\int_{\mathbb{R}^n} \exp\left(-q(y) \right)^{1/\delta} dy \right]^{\delta} \right) \tag{86d}$$

$$= \frac{\delta n}{2} \ln(2\pi\delta t) - \ln \left(\| \exp(-q_t) \|_{L^{1/\delta}(\mathbb{R}^n)} \right). \tag{86e}$$

As $\delta \to 0^+$, the first term vanishes via L'Hôpital's rule and the L^p norm becomes an L^{∞} norm, *i.e.*

$$u^{0^+}(x,t) = 0 - \ln \left(\| \exp\left(-q_t \right) \|_{L^{\infty}(\mathbb{R}^n)} \right) \tag{87a}$$

$$= 0 - \ln \left(\sup_{y \in \mathbb{R}^n} \exp\left(-q_t(y) \right) \right) \tag{87b}$$

$$= 0 - \ln \left(\exp\left(- \inf_{y \in \mathbb{R}^n} q_t(y) \right) \right) \tag{87c}$$

$$= \inf_{y \in \mathbb{R}^n} f(y) + \frac{\|x-y\|^2}{2t}, \tag{87d}$$

which is the desired limit.