A Hamilton-Jacobi-based Proximal Operator

Stanley Osher[a,1], **Howard Heaton**[b,1], **and Samy Wu Fung**[c,1,2]

[a]Dept. of Mathematics, University of California, Los Angeles; [b]Typal Research, Typal LLC; [c]Dept. of Applied Mathematics and Statistics, Colorado School of Mines

November 23, 2022

First-order optimization algorithms are widely used today. Two standard building blocks in these algorithms are proximal operators (proximals) and gradients. Although gradients can be computed for a wide array of functions, explicit proximal formulas are only known for limited classes of functions. We provide an algorithm, HJ-Prox, for accurately approximating such proximals. This is derived from a collection of relations between proximals, Moreau envelopes, Hamilton-Jacobi (HJ) equations, heat equations, and importance sampling. In particular, HJ-Prox smoothly approximates the Moreau envelope and its gradient. The smoothness can be adjusted to act as a denoiser. Our approach applies even when functions are only accessible by (possibly noisy) blackbox samples. We show HJ-Prox is effective numerically via several examples.

Proximal | Operator | Hamilton-Jacobi | Moreau | Optimization | Resolvent | Zeroth-Order | Importance Sampling | Cole-Hopf | Heat Equation

The rise of computational power and availability of big data brought great interest to first-order optimization methods. Second-order methods (*e.g.* Newton's method) are effective with moderately sized problems, but generally do not scale well due to memory requirements increasing quadratically with problem size and computation costs increasing cubically. First-order methods are often comprised of gradient and proximal operations, which are typically cheap to evaluate relative to problem size. Although gradients can be computed for many functions (or numerically approximated), the computation of proximals involves solving a small optimization problem. In special cases (*e.g.* with ℓ_1 norms), these subproblems admit closed-form solutions that can be quickly evaluated (*e.g.* see (1)). These formulas yield great utility in many applications. However, we are presently interested in the class of problems with (potentially nondifferentiable) objectives for which *proximal formulas are unavailable*.

We propose a new approach to compute proximal operators and corresponding Moreau envelopes for functions f. We leverage the fact that the Moreau envelope of f is the solution to a Hamilton-Jacobi (HJ) equation (2). The core idea is to add artificial viscosity to HJ equations and obtain explicit formulas for the proximal and Moreau envelopes using Cole-Hopf transformation (2, Sec. 4.5.2). This approach enables proximals and Moreau envelopes of arbitrary f to be approximated. Our proposed proximal approximations (called HJ-Prox) are computed *using only function evaluations* and can, thus, be used in a zeroth-order fashion when integrated within an optimization algorithm. Finally, an importance sampling procedure is employed to mitigate the curse of dimensionality when estimating the HJ-Prox in dimensions higher than three. Numerical experiments show HJ-Prox is effective when employed within optimization algorithms when the proximal is unavailable and for blackbox oracles. Our work can generally be applied to first-order proximal-based algorithms such as Alternating Direction Method of Multipliers (ADMM) and its variants (3–6), and operator splitting algorithms (7–11).

Proximal Operators and Moreau Envelopes

Consider a function $f\colon \mathbb{R}^n \to \mathbb{R}$ and time $t > 0$. The proximal prox_{tf} and the Moreau envelope u of f (12, 13) are defined by

$$\text{prox}_{tf}(x) \triangleq \underset{z\in\mathbb{R}^n}{\text{argmin}}\, f(z) + \frac{1}{2t}\|z-x\|^2 \qquad [1]$$

and

$$u(x,t) \triangleq \min_{z\in\mathbb{R}^n} f(z) + \frac{1}{2t}\|z-x\|^2. \qquad [2]$$

The proximal is the set of minimizers defining the envelope. As shown in Figure 1, the envelope u widens valleys of f while sharing global minimizers. A well-known result (*e.g.* see (1, 14)) states, if the envelope u is differentiable at x, then

$$\nabla u(x,t) = \frac{x - \text{prox}_{tf}(x)}{t}. \qquad [3]$$

Rearranging reveals

$$\text{prox}_{tf}(x) = x - t\nabla u(x,t). \qquad [4]$$

A key idea we use is to estimate the proximal by replacing u with a smooth approximation $u^\delta \in C^\infty(\mathbb{R})$, derived from a Hamilton-Jacobi (HJ) equation.

Hamilton-Jacobi Connection

The envelope u is a special case of the Hopf-Lax formula (2). Fix any $T > 0$. For all $t \in [0,T]$, the envelope u is a viscocity solution (*e.g.* see (15, Theorem 3.2)) to the HJ equation

$$\begin{cases} u_t + \frac{1}{2}\|\nabla u\|^2 = 0 & \text{in } \mathbb{R}^n \times (0,T] \\ u = f & \text{on } \mathbb{R}^n \times \{t=0\}. \end{cases} \qquad [5]$$

Fixing $\delta > 0$, the associated viscous HJ equation is

$$\begin{cases} u^\delta_t + \frac{1}{2}\|\nabla u^\delta\|^2 = \frac{\delta}{2}\Delta u^\delta & \text{in } \mathbb{R}^n \times (0,T] \\ u^\delta = f & \text{on } \mathbb{R}^n \times \{t=0\}. \end{cases} \qquad [6]$$

If f is bounded and Lipschitz, Crandall and Lions (16) show u^δ approximates u, *i.e.* $u^\delta \to u$ uniformly as $\delta \to 0^+$.

Contribution

Many objectives do not admit explicit proximal formulas (*e.g.* when objectives are either nonconvex or only accessible via an oracle). Yet, only using (possibly noisy) objective samples, we give a formula for accurately approximating such proximals.

Code is available at github.com/mines-opt-ml/hj-prox

[1]SO (Stanley Osher), HH (Howard Heaton), and SWF (Samy Wu Fung) contributed equally.

[2]Correspondence should be addressed to SWF via email: swufung@mines.edu.

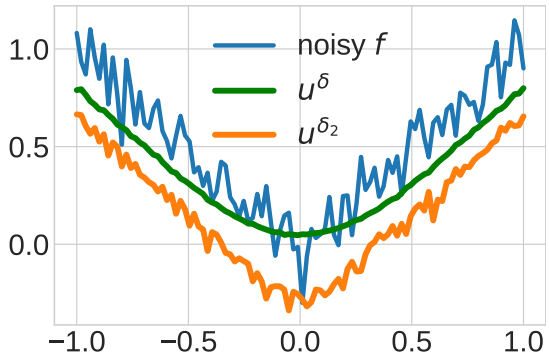


Fig. 1. Moreau envelope approximation u^δ using only noisy function samples with $\delta = 0.1$ and $\delta_2 = 0.01$.

Cole-Hopf Transformation

Using the transformation $v^\delta \triangleq \exp(-u^\delta/\delta)$, originally attributed to Cole and Hopf (2, 17), the function v^δ solves the heat equation, *i.e.*

$$\begin{cases} v_t^\delta - \frac{\delta}{2}\Delta v^\delta = 0 & \text{in } \mathbb{R}^n \times (0,T] \\ v^\delta = \exp(-f/\delta) & \text{on } \mathbb{R}^n \times \{t=0\}. \end{cases} \quad [7]$$

This transformation is of interest since v^δ can be expressed via the convolution formula (*e.g.* see (2) for a derivation)

$$v^\delta(x,t) = \Big(\Phi_{\delta t} * \exp(-f/\delta)\Big)(x) \quad [8a]$$

$$= (2\pi\delta t)^{-n/2} \int_{\mathbb{R}^n} \Phi_{\delta t}(x-y) \exp\left(-f(y)/\delta\right) \mathrm{d}y, \quad [8b]$$

where $\Phi_{\delta t}$ is a fundamental solution to [7], *i.e.*

$$\Phi_{\delta t}(x) \triangleq \begin{cases} (2\pi\delta t)^{-n/2} \exp\left(-|x|^2/(2\delta t)\right) & \text{in } \mathbb{R}^n \times (0,\infty) \\ 0 & \text{otherwise.} \end{cases} \quad [9]$$

Using algebraic manipulations, we recover the viscous solution

$$u^\delta(x,t) = -\delta \ln\Big(\Phi_{\delta t} * \exp(-f/\delta)\Big)(x) \quad \text{in } \mathbb{R}^n \times (0,T]. \quad [10]$$

Differentiating reveals

$$\nabla u^\delta(x,t) = -\delta \cdot \nabla\left[\ln\left(v^\delta(x,t)\right)\right] = -\delta \cdot \frac{\nabla v^\delta(x,t)}{v^\delta(x,t)}. \quad [11]$$

Importance Sampling

At first glance, the integral formula for v^δ in [11] may appear to require use of a grid for numerical estimation (and similarly for ∇v^δ). However, we may avoid such grids by noting v^δ can be written as an expectation, *i.e.*

$$v^\delta(x,t) = \Big(\Phi_{\delta t} * \exp(-f/\delta)\Big)(x) \quad [12a]$$

$$= \mathbb{E}_{y\sim\mathcal{N}(x,\delta t)}\left[\exp\left(-f(y)/\delta\right)\right], \quad [12b]$$

where $y \sim \mathcal{N}(x,\delta t)$ denotes $y \in \mathbb{R}^n$ is sampled from a normal distribution with mean x and standard deviation $\sqrt{\delta t}$. In practice, finitely many samples $y^i \sim \mathcal{N}(x,\delta t)$ are used to estimate [12b]. This can greatly reduce sampling complexity (18, 19). Differentiating v^δ with respect to x reveals

$$\nabla v^\delta(x,t) = -\frac{1}{\delta t} \cdot \mathbb{E}_{y\sim\mathcal{N}(x,\delta t)}\left[(x-y)\exp\left(-f(y)/\delta\right)\right]. \quad [13]$$

Algorithm 1 HJ-Prox – Approximation of Proximal Operator

1: HJ-Prox($x,\ t;\ f,\ \delta,\ N,\ \alpha,\ \varepsilon$) :
2: **for** $i \in [N]$:
3: Sample $y^i \sim \mathcal{N}(x, \delta t/\alpha)$
4: $z_i \leftarrow f(y^i)$
5: **if** $z_i < 0$:
6: **return** HJ-Prox($x,\ t;\ f + z_i + \varepsilon,\ \delta,\ N,\ \alpha,\ \varepsilon$)
7: **if** $\exp\left(-\alpha z_i/\delta\right) \le \varepsilon$:
8: **return** HJ-Prox($x,\ t;\ f,\ \delta,\ N,\ \alpha/2,\ \varepsilon$)
9: prox $\leftarrow$ softmax$(-\alpha z/\delta)^\top [y^1 \cdots y^N]$
10: **return** prox

Plugging [12b] and [13] into [11] enables ∇u^δ to be written as

$$\nabla u^\delta(x,t) = \frac{1}{t} \cdot \left(x - \frac{\mathbb{E}_{y\sim\mathbb{P}_{x,\delta t}}\left[y \cdot \exp\left(-f(y)/\delta\right)\right]}{\mathbb{E}_{y\sim\mathbb{P}_{x,\delta t}}\left[\exp\left(-f(y)/\delta\right)\right]}\right). \quad [14]$$

The above relation was used in (20). Here we take a further step, combining [4] and [14] to get an HJ-based estimate:

$$\mathrm{prox}_{tf}(x) = x - t\nabla u(x,t) \quad [15a]$$

$$\approx x - t\nabla u^\delta(x,t) \quad [15b]$$

$$= \frac{\mathbb{E}_{y\sim\mathbb{P}_{x,\delta t}}\left[y \cdot \exp\left(-f(y)/\delta\right)\right]}{\mathbb{E}_{y\sim\mathbb{P}_{x,\delta t}}\left[\exp\left(-f(y)/\delta\right)\right]}. \quad [15c]$$

As shown below, importance sampling enables efficient approximation of proximals in high dimensions (*e.g.* see Figure 2). Moreover, [15] estimates proximals *only using function values*, making it apt for zeroth-order optimization.

Numerical Considerations

A possible numerical challenge in our formulation is to address numerical instabilities arising from the exponential term underflowing with limited numerical precision, due to either δ being small or $f(y)$ being large. To this end, note the proximal formula may equivalently be re-scaled via

$$\mathrm{prox}_{tf}(x) = \mathrm{prox}_{\frac{t}{\alpha}\alpha f}(x) \quad [16a]$$

$$\approx \frac{\mathbb{E}_{y\sim\mathbb{P}_{x,\delta t/\alpha}}\left[y \cdot \exp\left(-\alpha f(y)/\delta\right)\right]}{\mathbb{E}_{y\sim\mathbb{P}_{x,\delta t/\alpha}}\left[\exp\left(-\alpha f(y)/\delta\right)\right]}, \quad [16b]$$

where t is replaced by t/α and f by αf in [15].

In this case, if f/δ becomes too large with respect to numerical precision limitations, it may be scaled down with a corresponding α. To make the implementation stable, we check whether we obtain an underflow with $\exp(\alpha f(y)/\delta)$ and rescale α using a linesearch-like approach. In particular, we recursively halve α until $\exp(\alpha f(y)/\delta) > \varepsilon$ for a tolerance ε (see line 7 of Algorithm 1. Yet, small α makes the variance large and more samples may be required to accurately estimate the expectations. Another mitigation is to adaptively rescale f based on the number of recursive steps taken in HJ-Prox.

Large δ can be used to smooth approximations and mitigate the stochastic characteristics of HJ-Prox. Another potential instability that may arise is when f is negative in certain parts of the domain. In this case, $\exp(\alpha f(y)/\delta)$ may overflow. To remedy this, we check whether $f(y)$ is negative and recursively shift the function until it is nonnegative (see line 5 of Algorithm 1).

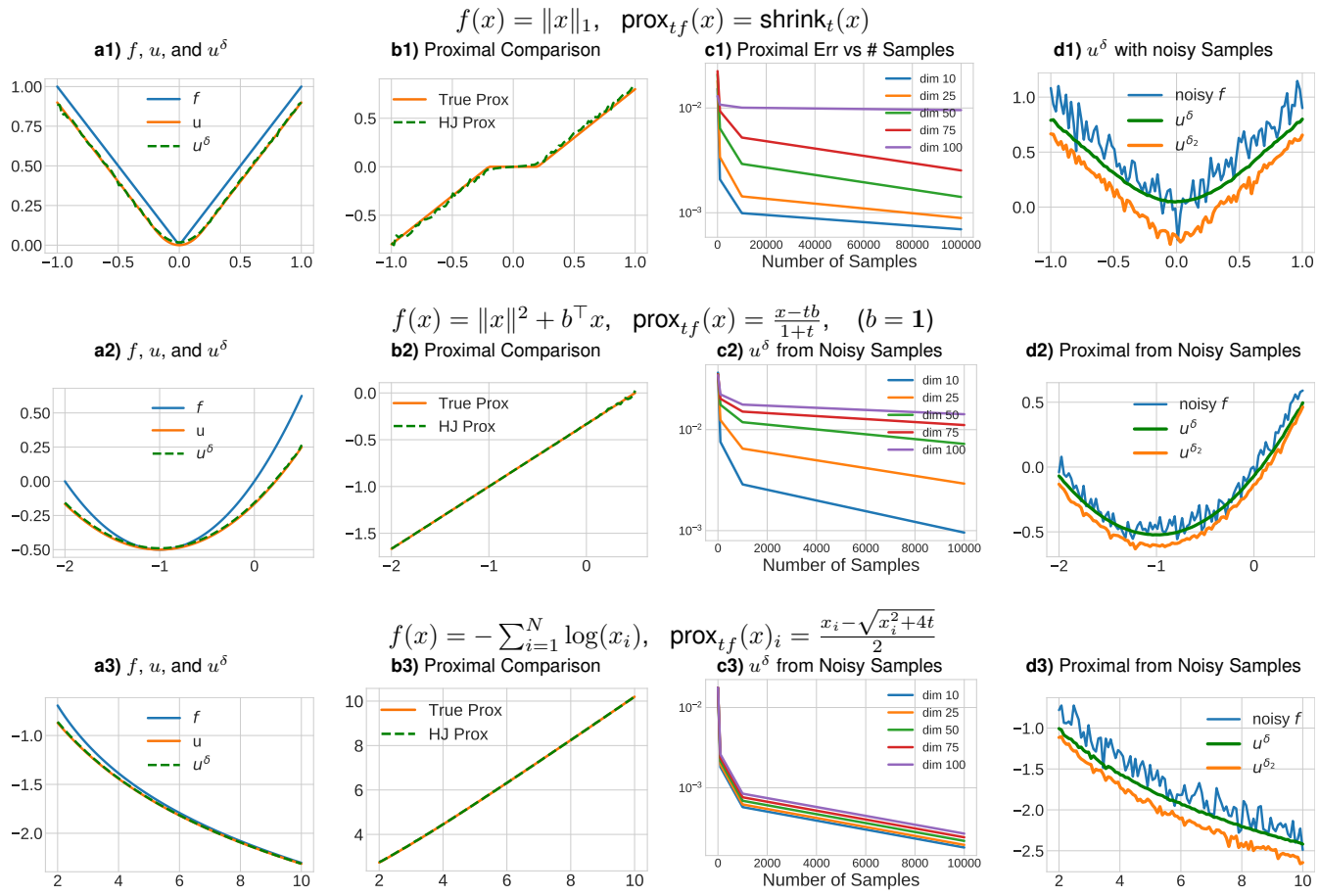


Fig. 2. (a1, a2, a3): Plots for function f, exact Moreau envelope u, and HJ-based Moreau envelope u^δ. **(b1, b2, b3):** Plots for true proximal and approximate HJ-based proximal operators. **(c1, c2, c3):** Proximal approximations across different dimensions and samples. **(d1, d2, d3):** HJ-based Moreau envelopes u^δ obtained from *noisy function samples*. Here, we use $\delta = 10^{-1}$ and $\delta_2 = 10^{-2}$. As expected, higher δ values have a stronger smoothing property. The HJ-proximals are good approximations of the true proximal operators (seen through the Moreau envelopes) and can even be applied when only (potentially noisy) samples are available. For the noisy case, we obtain a C^∞ approximation of the underlying function f. For these experiments, we use $t = 0.1, 0.5, 2.0$ for rows 1, 2, and 3, respectively.

Convergence Analysis

The arguments above give intuition for a proximal approximation. However, having now the formula [15], we may formalize its utility without reference to differential equations. Below we define two standard classes of functions used in optimization.

Definition 1 (Weakly Convex). *For $\rho > 0$, a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is ρ-weakly convex if $f(x) + \frac{\rho}{2}\|x\|^2$ is convex.*

Definition 2 (L-Smooth). *For $L > 0$, a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is L-smooth if its gradient ∇f exists and is L-Lipschitz.*

Our main result shows HJ-Prox converges to the proximal.

Theorem 1 (Proximal Approximation). *If $f\colon \mathbb{R}^n \to \mathbb{R}$ is ρ-weakly convex, for some $\rho > 0$, and either L-smooth or L-Lipschitz, then, for all $x \in \mathbb{R}^n$, and $t \in (0, 1/\rho)$, the proximal $\text{prox}_{tf}(x)$ is unique and, if $u(x,t) \geq 0$, then*

$$\lim_{\delta \to 0^+} \frac{\mathbb{E}_{y\sim\mathcal{N}(x,\delta t/)}\left[y \cdot \exp\left(-f(y)/\delta\right)\right]}{\mathbb{E}_{y\sim\mathcal{N}(x,\delta t/)}\left[\exp\left(-f(y)/\delta\right)\right]} = \text{prox}_{tf}(x). \qquad [17]$$

Remark 1 (Smoothing Property). *In practice, we must pick positive δ. Thankfully, increasing δ comes with the benefit of smoothing estimates (due to the Laplacian in the viscous HJ equation), as shown in rightmost column of Figure 2.*

Related Works

Our proposal closely relates to zeroth-order optimization algorithms, which do not require gradients. In fact, HJ-Prox does not require differentiability of f. Related methods include Random Gradients (21–24), sparsity-based methods (25–27), derivative-free quasi-Newton methods (28–30), finite-difference-based methods (31, 32), numerical quadrature-based methods (33, 34), Bayesian methods (29), and comparison methods (35). As proximals closely relate to gradient of Moreau envelopes, our work relates to methods that minimize Moreau envelopes (or their approximations) (17, 20, 36–40).

The theoretical results in our work is closely related to the study of asymptotics as $\delta \to 0$ of integrals containing expressions of the form $\exp(-f/\delta)$, *i.e.* Laplace's method (2). Moreover, the idea of adding artificial diffusion to Burgers' equation and then applying Cole-Hopf transformation to approximate the gradient of the solution to the HJ equation has been largely developed in (2) in the context of obtaining solutions to conservation laws in 1D. The connections between Hopf-Lax and Cole-Hopf was first introduced in the context of machine learning in (17) and in the context of global optimization in (20).

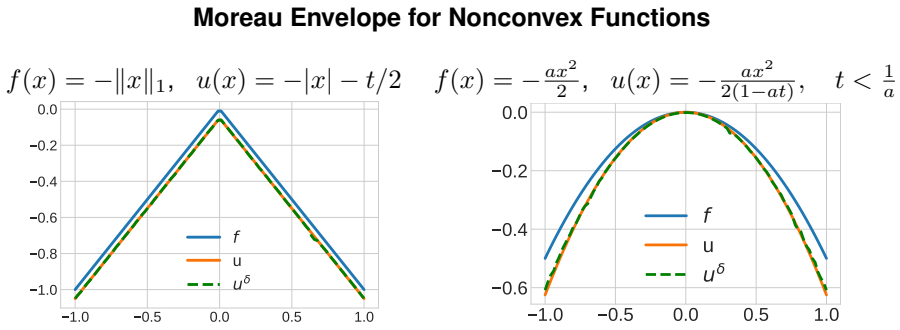


Fig. 3. HJ-based Moreau envelope for nonconvex functions with $t = 0.1$ and $t = 0.2$ in the left and right figures, respectively.

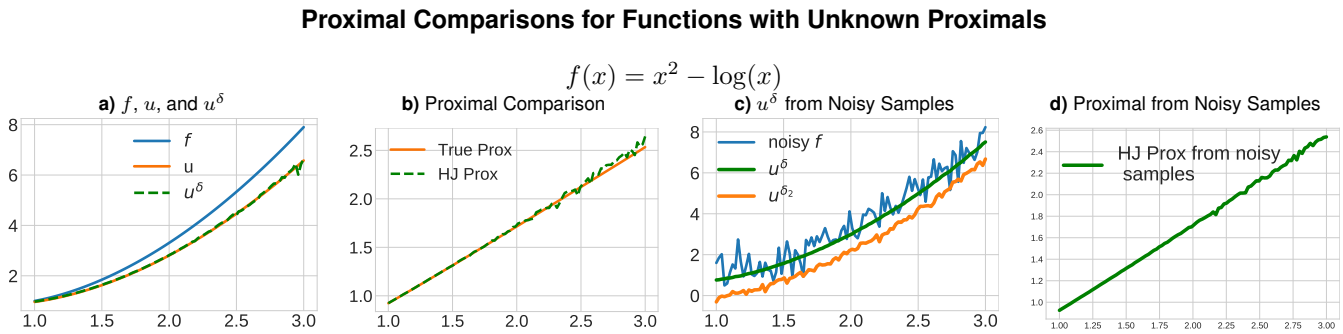


Fig. 4. (a): Plots for function f, exact Moreau envelope u, and HJ-based Moreau envelope u^δ. **(b):** Plots for true proximal and approximate HJ-based proximal operators. **(c):** HJ-based Moreau envelopes u^δ obtained from *noisy function samples*. **(d):** HJ-based proximal computed *using noisy function samples*. Since there is no analytic proximal formula, we obtain the "true" proximal by solving the optimization Eq. (1) using gradient descent. The HJ-based proximal is a good approximation of the true proximal operators and can even be applied when only (potentially noisy) samples are available. As in the analytic case, we obtain a C^∞ approximation of the underlying function f in the noisy case. Here, $\delta = 0.1$ for the noiseless case and $\delta = 0.5$ and $\delta_2 = 0.1$ for the noisy case.

Numerical Experiments

Examples herein show HJ-Prox (Algorithm 1) can

- approximate proximals *and* smooth noisy samples,
- converge comparably to existing algorithms, and
- solve a new class of zeroth-order optimization problems.

Each item is addressed by a set of experiments. Regarding the last item, to our knowledge, HJ-Prox is the first tool to enable faithful solution estimation for constrained problems where the objective is only accessible via noisy blackbox samples.

Proximal and Moreau Envelope Estimation. Herein we compare HJ-Prox to known proximal operators. Figure 2 shows HJ-Prox for three functions (absolute value, quadratic, and log barrier) whose proximals are known. In the leftmost column (a), we show the Moreau envelope $u(x,t)$ given by [2], and an estimate of Moreau envelope using the HJ-Prox $u^\delta(x,t)$. Given the close connection between proximals and Moreau envelopes, we believe this visual is a natural and intuitive way to gauge whether the proximal operator is accurate. Column (b) juxtaposes the true proximal and HJ-Prox. Column (c) shows the accuracy of HJ-Prox across different dimensions and numbers of samples. In the rightmost column (d), we estimate Moreau envelopes using HJ-Prox *using noisy function values*. The resulting envelopes are smooth since u^δ is a smooth (*i.e.* C^∞) approximation of u. Thus, HJ-Prox can be used to obtain smooth estimates from noisy observations.

Figure 3 shows Moreau envelopes for nonconvex functions f. As in the other example, here HJ-based Moreau envelope estimates also accurately approximate Moreau envelopes. Note these proximals may be well-defined only for small time t (as the proximal operator objective in [1] is strongly convex for small t). Lastly, we apply HJ-Prox with a function that has *no analytic formula* for its proximal or Moreau envelope in Figure 4. In this experiment, we obtain a "true" Moreau envelope and proximal operator by solving the minimization problem [1] iteratively via gradient descent. Faithful recovery is shown in Figures 4a and 4b, and smoothing in Figure 4c.

Optimization with Proximable Function. This experiment juxtaposes HJ-prox and an analytic proximal formula in an optimization algorithm. Consider the Lasso problem (41)

$$\min_{x\in\mathbb{R}^{1000}} \frac{1}{2}\|Ax-b\|_2^2 + 0.1\|x\|_1, \qquad [18]$$

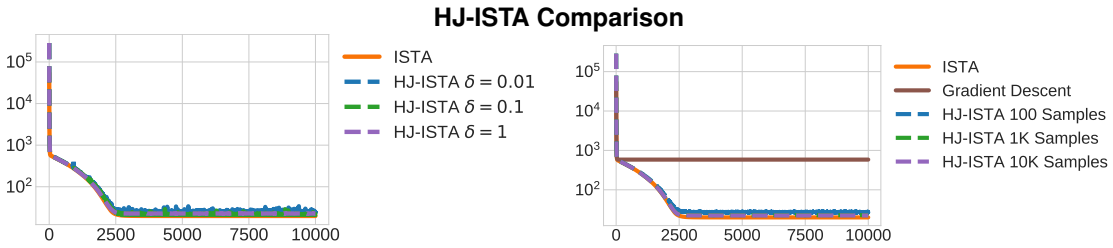
where entries of $A \in \mathbb{R}^{500\times 1000}$ and $b \in \mathbb{R}^{500}$ are i.i.d. Gaussian samples. The iterative soft thresholding algorithm (ISTA) (42) defines a sequence of solution estimates $\{x^k\}$ for all $k \in \mathbb{N}$ via

$$x^{k+1} = \text{shrink}\left(x^k - \beta A^\top (Ax^k - b);\ 0.01\beta\right), \qquad [19]$$

where the shrink operator defined element-wise by

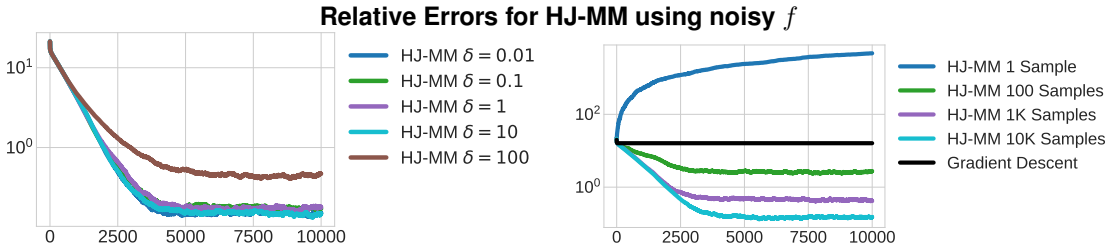
$$\text{shrink}(x;\ t) \triangleq \text{sign}(x)\max(0,\ |x| - t). \qquad [20]$$

Figure 5 compares the convergence of ISTA using the shrink operator in [20] and HJ-Prox estimates of the shrink. To ensure convergence, we choose $\beta = 1/\|A^\top A\|_2$. Our experiments show HJ-based ISTA can solve Lasso, up to an error tolerance.



(a) Varying Smoothing δ, Fixed # Samples $N = 1000$ **(b)** Varying # Samples N, Fixed Smoothing $\delta = 1.0$

Fig. 5. Convergence plots showing function value for solution estimates $\{x^k\}$ when solving the LASSO problem [18] with ISTA, juxtaposing use of an analytic proximal formula, gradient descent (*i.e.* ignoring the proximal), and the approximate HJ-prox (Algorithm 1). Plots with HJ-prox show averaged results from 30 trials with distinct random seeds. To ensure the proximal is playing a role in the optimization process, we also show a function value history of gradient descent applied to the unregularized least squares problem in [18] (*i.e.* , with no ℓ_1 norm term).



(a) Varying Smoothing δ, Fixed # Samples N =10K **(b)** Varying # Samples N, Fixed Smoothing $\delta = 10$

Fig. 6. Convergence plots showing relative errors for solution estimates $\{x^k\}$ when solving the minimization problem [21] with linearized method of multipliers and HJ-prox (Algorithm 1). Each plot shows averaged results from 30 trials with distinct random seeds. Due to the noise , we observe in (a) that a larger $\delta = 10$ leads to a better approximation, but too large ($\delta = 100$) leads to oversmoothing and reduces accuracy. We find $\delta = 10$ to be most optimal, and (b) shows that more samples lead to more accurate approximations (as expected). To ensure the proximal is playing a role in the optimization process, we also show the relative error when gradient decent is applied to the constraint residual in [21] (*i.e.* , we only minimize constraint residual). Indeed, gradient descent performs poorly by comparison.

Optimization with Noisy Objective Oracles. Consider a constrained minimization problem where objective values f can only be accessed via a noisy oracle* $\mathcal{O}$. Our task is to solve

$$\min_{x\in\mathbb{R}^{1000}} \mathbb{E}[\mathcal{O}(x)] \quad \text{s.t.} \quad Ax = b, \tag{21}$$

where A and b are as in the prior experiment and the expectation $\mathbb{E}$ is over oracle noise. To model "difficult" settings (*e.g.* when a singular value decomposition of A is unavailable), we do *not* use any projections onto the feasible set. As knowledge of the structure of $\mathcal{O}$ is *unknown* to the solver, we emphasize schemes for solving [21] must use zeroth-order optimization schemes (29). Here, each oracle call returns

$$\mathcal{O}(x) = (1+\varepsilon)\cdot\|Wx\|_1, \quad \text{where } \varepsilon \sim \mathcal{N}(0,\sigma^2), \tag{22}$$

with a *new noise sample* $\varepsilon \in \mathbb{R}$ used in each oracle evaluation, $\sigma = 0.005$, and $W \in \mathbb{R}^{1000\times 1000}$ a fixed Gaussian matrix. In words, the noise has magnitude 0.5% of $\|Wx\|_1$. Although the oracle structure is shown by [22], our task is to solve [21] *without* such knowledge. We do this with the linearized method of multipliers (*e.g.* see Section 3.5 in (9)). Specifically, for each index $k \in \mathbb{N}$, the update formulas for the solution estimates $\{x^k\}$ and corresponding dual variables $\{u^k\}$ are

$$x^{k+1} = \text{prox}_{t\mathcal{O}}\left(x^k - tA^\top(u^k + \lambda(Ax^k - b))\right) \tag{23a}$$

$$u^{k+1} = u^k + \lambda(Ax^{k+1} - b), \tag{23b}$$

with step sizes $t = 1/\|A^\top A\|_2$ and $\lambda = 1/2$. Without noise ε, convergence occurs if $t\lambda\|A^\top A\|_2 < 1$ (9), justifying our choices for t and λ. The proximal $\text{prox}_{t\mathcal{O}}$ is estimated by HJ-prox.

*Here $\mathcal{O}$ is a noisy function, *not* to be confused with "Big O" often used to describe limit behaviors.

We *separately* solve the optimization problem using full knowledge of the objective $\|Wx\|_1$ without noise; doing this enables us to plot the relative error of the sequence $\{x^k\}$ in Figure 6. All the plots show $\{x^k\}$ converges to the optimal $x^\star$, up to an error threshold, regardless of the choice of δ and number of samples N. Notice Figure 6a shows "small" values of δ give comparable accuracy, but that oversmoothing with "large" $\delta = 100$ degrades performance of the algorithm. These plots also illustrate the HJ-prox formula is efficient with respect to calls to the oracle $\mathcal{O}$. Indeed, note the plots in Figure 6b that decrease relative error use, at each iteration, respectively use 0.1, 1, and 10 oracle calls per dimension of the problem! We hypothesize the smoothing effect of the viscous u^δ and averaging effect of importance sampling contribute to the observed convergence. In this experiment, HJ-prox converges to within an error tolerance, is efficient with respect to oracle calls, and smooths Gaussian noise.

Conclusion

We propose a novel algorithm, HJ-prox, for efficiently approximating proximal operators. This is derived from approximating Moreau envelopes via viscocity solutions to Hamilton-Jacobi (HJ) equations, as given via the Hopf-Lax formula. Upon rewriting this approximation in terms of expectations, we use importance sampling to avoid discretizing the integrals, thereby mitigating the curse of dimensionality. Our numerical examples show HJ-Prox is effective for a collection of functions, both with and without known proximal formulas. Moreover, HJ-prox can be effectively used in constrained optimization problems *even when only noisy objective values are available.*

Acknowledgements

SO thanks the funding from AFOSR MURI FA9550-18-1-0502, ONR:N00014-20-1-2093 and N00014-20-1-2787, and NSF DMS 2208272 and 1952339.

References

1. A Beck, *First-order methods in optimization*. (SIAM), (2017).
2. LC Evans, Partial Differential Equations. *Graduate Stud. Math.* **19** (2010).
3. MJ Powell, A method for nonlinear constraints in minimization problems. *Optimization* pp. 283–298 (1969).
4. S Boyd, et al., Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations Trends Mach. learning* **3**, 1–122 (2011).
5. MR Hestenes, Multiplier and gradient methods. *J. optimization theory applications* **4**, 303–320 (1969).
6. SW FUNG, S TYRVÄINEN, L RUTHOTTO, E HABER, Admm-softmax: An admm approach for multinomial logistic regression. *Electron. Transactions on Numer. Analysis* **52**, 214–229 (2020).
7. J Eckstein, DP Bertsekas, On the douglas—rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Program.* **55**, 293–318 (1992).
8. PL Lions, B Mercier, Splitting algorithms for the sum of two nonlinear operators. *SIAM J. on Numer. Analysis* **16**, 964–979 (1979).
9. EK Ryu, W Yin, *Large-Scale Convex Optimization*. (Cambridge University Press), (2022).
10. D Davis, W Yin, A three-operator splitting scheme and its optimization applications. *Set-valued variational analysis* **25**, 829–858 (2017).
11. T Goldstein, S Osher, The split bregman method for l1-regularized problems. *SIAM journal on imaging sciences* **2**, 323–343 (2009).
12. JJ 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).
13. HH Bauschke, PL Combettes, , et al., *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. (Springer), 2nd edition, (2017).
14. RT Rockafellar, *Convex Analysis*. (Princeton University Press) Vol. 18, (1970).
15. LC Evans, Envelopes and nonconvex Hamilton–Jacobi equations. *Calc. Var. Partial. Differ. Equations* **50**, 257–282 (2014).
16. MG Crandall, PL Lions, Two approximations of solutions of Hamilton-Jacobi equations. *Math. computation* **43**, 1–19 (1984).
17. P Chaudhari, A Oberman, S Osher, S Soatto, G Carlier, Deep relaxation: partial differential equations for optimizing deep neural networks. *Res. Math. Sci.* **5**, 1–30 (2018).
18. T Kloek, HK Van Dijk, Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econom. J. Econom. Soc.* pp. 1–19 (1978).
19. ST Tokdar, RE Kass, Importance sampling: a review. *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 54–60 (2010).
20. H Heaton, SW Fung, S Osher, Global solutions to nonconvex problems by evolution of hamilton-jacobi pdes. *arXiv preprint arXiv:2202.11014* (2022).
21. YM Ermoliev, RB Wets, *Numerical techniques for stochastic optimization*. (Springer-Verlag), (1988).
22. D Kozak, S Becker, A Doostan, L Tenorio, Stochastic subspace descent. *arXiv preprint arXiv:1904.01145* (2019).
23. D Kozak, S Becker, A Doostan, L Tenorio, A stochastic subspace approach to gradient-free optimization in high dimensions. *Comput. Optim. Appl.* **79**, 339–368 (2021).
24. D Kozak, C Molinari, L Rosasco, L Tenorio, S Villa, Zeroth order optimization with orthogonal random directions. *arXiv preprint arXiv:2107.03941* (2021).
25. H Cai, D Mckenzie, W Yin, Z Zhang, Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling. *SIAM J. on Optim.* **32**, 687–714 (2022).
26. H Cai, Y Lou, D McKenzie, W Yin, A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization in *International Conference on Machine Learning*. (PMLR), pp. 1193–1203 (2021).
27. I Slavin, D McKenzie, Adapting zeroth order algorithms for comparison-based optimization. *arXiv preprint arXiv:2210.05824* (2022).
28. AS Berahas, RH Byrd, J Nocedal, Derivative-free optimization of noisy functions via quasi-newton methods. *SIAM J. on Optim.* **29**, 965–993 (2019).
29. J Larson, M Menickelly, SM Wild, Derivative-free optimization methods. *Acta Numer.* **28**, 287–404 (2019).
30. J Moré, S Wild, Benchmarking derivative-free optimization algorithms. *SIAM J. on Optim.* **20**, 172–191 (2009).
31. HJM Shi, MQ Xuan, F Oztoprak, J Nocedal, On the numerical performance of derivative-free optimization methods based on finite-difference approximations. *arXiv preprint arXiv:2102.09762* (2021).
32. HJM Shi, Y Xie, MQ Xuan, J Nocedal, Adaptive finite-difference interval estimation for noisy derivative-free optimization. *arXiv preprint arXiv:2110.06380* (2021).
33. B Kim, H Cai, D McKenzie, W Yin, Curvature-aware derivative-free optimization. *arXiv preprint arXiv:2109.13391* (2021).
34. LB Almeida, A learning rule for asynchronous perceptrons with feedback in a combinatorial environment in *Artificial neural networks: concept learning*. pp. 102–111 (1990).
35. H Cai, D Mckenzie, W Yin, Z Zhang, A one-bit, comparison-based gradient estimator. *Appl. Comput. Harmon. Analysis* **60**, 242–266 (2022).
36. P Chaudhari, et al., Entropy-sgd: Biasing gradient descent into wide valleys. *J. Stat. Mech. Theory Exp.* **2019**, 124018 (2019).
37. K Scaman, L Dos Santos, M Barlier, I Colin, A simple and efficient smoothing method for faster optimization and local exploration. *Adv. Neural Inf. Process. Syst.* **33**, 6503–6513 (2020).
38. D Davis, D Drusvyatskiy, Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988* (2018).
39. D Davis, D Drusvyatskiy, Stochastic model-based minimization of weakly convex functions. *SIAM J. on Optim.* **29**, 207–239 (2019).
40. D Davis, M Díaz, D Drusvyatskiy, Escaping strict saddle points of the Moreau envelope in nonsmooth optimization. *SIAM J. on Optim.* **32**, 1958–1983 (2022).
41. I Daubechies, M Defrise, C De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. on Pure Appl. Math. A J. Issued by Courant Inst. Math. Sci.* **57**, 1413–1457 (2004).
42. A Beck, M Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* **2**, 183–202 (2009).
43. W Rudin, *Real and Complex Analysis*. (McGraw-Hill), (1966).

Proofs

For concise expression below, for $t > 0$ and $\delta > 0$ we define

$$\phi_t(z) \triangleq f(z) + \frac{1}{2t}\|z - x\|^2, \tag{24}$$

$\phi_t^\star \triangleq \inf\{\phi_t(y) : y \in \mathbb{R}^n\}$, and

$$\sigma_\delta(z) \triangleq \frac{\exp\left(-\phi(z)/\delta\right)}{\|\exp(-\phi/\delta)\|_{L^1(\mathbb{R}^n)}}. \tag{25}$$

Lemma 1. *If the conditions of Theorem 1 hold, then*

$$\int_{\mathbb{R}^n} \sigma_\delta(y)\ \mathrm{d}y = 1, \quad \sigma_\delta(y) \geq 0, \quad \text{for all } y \in \mathbb{R}^n, \tag{26}$$

and if $r \in (0, 1)$, then for all polynomials p of positive degree

$$\lim_{\delta \to 0^+} \int_{\mathbb{R}^n - \mathcal{B}(\xi^\star, r)} \sigma_\delta(y) p(\|y - \xi^\star\|)\ dy = 0. \tag{27}$$

Proof. By algebraic limit laws, it suffices to verify [27] for any $p(x) = x^k$ with $k \geq 1$, and we proceed as follows. First we show σ_δ satisfies properties to be a probability density (Step 1). We show various L^p norm limits hold for the numerator (Step 2) and denominator (Step 3) of integrating [27]. Combining these limits gives [27] (Step 4).

Step 1 The numerator and denominator in the definition [25] for σ_δ are nonnegative, making $\sigma_\delta \geq 0$ everywhere. By the choice of t, ϕ_t is $\theta \triangleq 1/t - \rho$ strongly convex, and so it admits a unique minimizer $\xi^\star = \mathrm{prox}_{tf}(x)$ and satisfies

$$\phi_t(y) \geq \phi_t^\star + \langle 0, y - \xi^\star\rangle + \frac{\theta}{2}\|y - \xi^\star\|^2, \quad \text{for all } y \in \mathbb{R}^n. \tag{28}$$

Consequently,

$$0 < e^{-\frac{\phi_t(y)}{\delta}} \leq e^{-\frac{\phi_t^\star + \frac{\theta}{2}\|y-\xi^\star\|^2}{\delta}}, \quad \text{for all } y \in \mathbb{R}^n. \tag{29}$$

Since the upper bound above is an exponential that decays quadratically, the middle term in [29] is integrable over $\mathbb{R}^n$. As $\phi_t^\star = u(\xi^\star, t) \geq 0$ by hypothesis, the denominator in the definition of σ_δ is positive. Then [26] readily follows.

Step 2 A classic result in analysis (*e.g.* see (43, Exercise 3.4)) states L^p norms converge to the L^∞ norm as $p \to \infty$, and so

$$\lim_{\delta \to 0^+} \left\|e^{-\phi_t}\right\|_{L^{\frac{1}{\delta}}(\mathbb{R}^n)} = \|e^{-\phi_t}\|_{L^\infty(\mathbb{R}^n)} = e^{-\phi_t^\star}, \tag{30}$$

where the $L^{\frac{1}{\delta}}$ norm is always finite by Step 1 and the final equality holds since the exponential is maximized by $\phi_t^\star$.

Step 3 Integrating the numerator of [27] for $p(x) = x^k$ gives

$$\int_{\mathbb{R}^n - \mathcal{B}(\xi^\star, r)} e^{-\frac{\phi_t(y)}{\delta}} \|y - \xi^\star\|^k\ \mathrm{d}y \tag{31a}$$

$$\leq \int_r^\infty e^{-\frac{\phi_t^\star + \frac{\theta\tau^2}{2}}{\delta}} \tau^k \cdot n|\mathcal{B}(\xi^\star, 1)|\tau^{n-1}\ \mathrm{d}\tau \tag{31b}$$

$$= n|\mathcal{B}(\xi^\star, 1)| \cdot \int_r^\infty e^{-\frac{\phi_t^\star + \frac{\theta\tau^2}{2} - (n+k-1)\ln(\tau^\delta)}{\delta}}\ \mathrm{d}\tau, \tag{31c}$$

where the first inequality follows from a change of variables to polar coordinates and using the strong convexity of ϕ_t in [28], and the final line by algebraic properties of logarithms.

Now define

$$\varepsilon \triangleq \frac{\theta}{4(n + k - 1)} > 0, \tag{32}$$

where the denominator is positive since $n \geq 1$ as p has positive degree. For all $0 < \delta < \varepsilon$, observe

$$\tau > 1 \implies \tau^\delta < \tau^\varepsilon \quad \text{and} \quad \tau \leq 1 \implies \tau^\delta \leq 1^\varepsilon, \tag{33}$$

i.e.

$$\tau^\delta \leq \max(\tau, 1)^\varepsilon, \quad \text{for all } \delta \in (0, \varepsilon). \tag{34}$$

Whence, continuing [31], we deduce, for all $\delta \in (0, \varepsilon)$,

$$\int_{\mathbb{R}^n - \mathcal{B}(\xi^\star, r)} e^{-\frac{\phi_t(y)}{\delta}} \|y - \xi^\star\|^k\ \mathrm{d}y \tag{35a}$$

$$\leq n|\mathcal{B}(\xi^\star, 1)| \cdot \int_r^\infty e^{-\frac{\phi_t^\star + \frac{\theta\tau^2}{2} - \varepsilon(n+k-1)\ln(\max(\tau,1))}{\delta}}\ \mathrm{d}\tau. \tag{35b}$$

Let $q(y)$ be the numerator inside the exponential in the integrand. Taking the limit

$$\lim_{\delta \to 0^+} \|e^{-q}\|_{L^{\frac{1}{\delta}}([r,\infty))} = \left\|e^{-q}\right\|_{L^\infty([r,\infty))}. \tag{36}$$

Let $\tau^\star$ be the minimizer of q over $[r, \infty)$. If $\tau^\star > 1$, then the first order necessary condition implies, together with [32],

$$0 = \theta\tau^\star - \frac{\varepsilon(n + k - 1)}{\tau^\star} \implies \tau^\star = \sqrt{\frac{\varepsilon(n + k - 1)}{\theta}} = \frac{1}{2}, \tag{37}$$

a contradiction. Consequently, $\tau^\star \leq 1$. Since q is quadratic in τ and strictly increasing on $[r, 1)$, we deduce $\tau^\star = r$. Thus,

$$\|e^{-q}\|_{L^\infty([r,\infty))} = e^{-\phi_t^\star - \frac{\theta r^2}{2}}. \tag{38}$$

Furthermore, note

$$\lim_{\delta \to 0^+} [n|\mathcal{B}(\xi^\star, 1)|]^\delta = 1. \tag{39}$$

Together [35], [38], and [39] imply

$$\lim_{\delta \to 0^+} \left[\int_{\mathbb{R}^n - \mathcal{B}(\xi^\star, r)} e^{-\frac{\phi_t(y)}{\delta}} \|y - \xi^\star\|^k\ \mathrm{d}y\right]^\delta \leq e^{-\phi_t^\star - \frac{\theta r^2}{2}}. \tag{40}$$

Step 4 By [30] and [40] and the definition of σ_δ,

$$\lim_{\delta \to 0^+} \left[\int_{\mathbb{R}^n - \mathcal{B}(\xi^\star, r)} \sigma_\delta(y)\|y - \xi^\star\|^k\ \mathrm{d}y\right]^\delta \leq \underbrace{\frac{e^{-\phi_t^\star - \frac{\theta r^2}{2}}}{e^{-\phi_t^\star}}}_{\triangleq \gamma} < 1. \tag{41}$$

Consequently, there is $\bar{\delta} > 0$ such that

$$\left[\int_{\mathbb{R}^n - \mathcal{B}(\xi^\star, r)} \sigma_\delta(y)\|y - \xi^\star\|^k\ \mathrm{d}y\right]^\delta \leq \frac{\gamma + 1}{2}, \quad \text{for all } \delta \in (0, \bar{\delta}], \tag{42}$$

where we note $(\gamma + 1)/2 \in (\gamma, 1)$, and so

$$\lim_{\delta \to 0^+} \int_{\mathcal{S}} \sigma_\delta(y)\|y - \xi^\star\|^k\ \mathrm{d}y \leq \lim_{\delta \to 0^+} \left(\frac{\gamma + 1}{2}\right)^{1/\delta} = 0, \tag{43}$$

as desired. □

Below we restate and prove the main theorem, which is an extension of a lemma in Section 4.5.2 of (2).

Theorem 1 (Proximal Approximation). *If $f\colon \mathbb{R}^n \to \mathbb{R}$ is ρ-weakly convex, for some $\rho > 0$, and either L-smooth or L-Lipschitz, then, for all $x \in \mathbb{R}^n$, and $t \in (0, 1/\rho)$, the proximal $\operatorname{prox}_{tf}(x)$ is unique and, if $u(x,t) \geq 0$, then*

$$\lim_{\delta \to 0^+} \frac{\mathbb{E}_{y\sim\mathcal{N}(x,\delta t/)}\left[y \cdot \exp\left(-f(y)/\delta\right)\right]}{\mathbb{E}_{y\sim\mathcal{N}(x,\delta t/)}\left[\exp\left(-f(y)/\delta\right)\right]} = \operatorname{prox}_{tf}(x). \tag{44}$$

Proof. Let $x \in \mathbb{R}^n$ and $t > 0$ be given. For notational compactness, denote the HJ-prox formula by

$$\xi^\delta \triangleq \frac{\mathbb{E}_{y\sim\mathcal{N}(x,\delta t)}\left[y \cdot \exp\left(-f(y)/\delta\right)\right]}{\mathbb{E}_{y\sim\mathcal{N}(x,\delta t)}\left[\exp\left(-f(y)/\delta\right)\right]}, \quad \text{for all } \delta > 0, \tag{45}$$

denote the proximal by $\xi^\star \triangleq \operatorname{prox}_{tf}(x)$, and note $\phi_t^\star = \phi_t(\xi^\star)$. As argued in Lemma 1, $\xi^\star$ is well-defined. We first bound $\phi_t - \phi_t^\star$ using Jensen's inequality (Step 1). Second, we show $\phi_t(\xi^\delta) \to \phi_t(\xi^\star)$ (Step 2). The strong convexity of ϕ_t enables us to establish the desired limit (Step 3).

Step 1 Note ξ^δ can be rewritten via

$$\xi^\delta = \left[\int_{\mathbb{R}^n} e^{-\frac{\phi(y)}{\delta}} \mathrm{d}y\right]^{-1} \int_{\mathbb{R}^n} y \cdot e^{-\frac{\phi_t(y)}{\delta}} \, \mathrm{d}y. \tag{46}$$

Using σ_δ, the estimate can be more concisely written via

$$\xi^\delta = \int_{\mathbb{R}^n} \sigma_\delta(y) y \, \mathrm{d}y = \mathbb{E}_{y\sim\mathbb{P}_{\sigma_\delta}}\left[y\right], \tag{47}$$

where the expectation holds by utilizing the fact [26] shows σ_δ defines a probability density. Thus, Jensen's inequality may be applied to reveal

$$0 \leq \phi_t^\star \leq \phi_t(\xi^\delta) = \phi_t\left(\mathbb{E}_{y\sim\sigma_\delta}[y]\right) \leq \mathbb{E}_{y\sim\sigma_\delta}\left[\phi_t(y)\right]. \tag{48}$$

In integral form, we may subtract $\phi_t^\star$ to write

$$0 \leq \phi_t(\xi^\delta) - \phi_t^\star \leq \int_{\mathbb{R}^n} \sigma_\delta(y)[\phi_t(y) - \phi_t^\star] \, \mathrm{d}y. \tag{49}$$

Step 2 Let $\varepsilon > 0$ be given. To deduce $\phi_t(\xi^\delta) \to \phi_t^\star$, we verify there is $\delta^\star > 0$ such that

$$|\phi_t(\xi^\delta) - \phi_t^\star| \leq \varepsilon, \quad \text{for all } \delta \in (0, \delta^\star]. \tag{50}$$

By [49], the relation [50] holds if there is such a $\delta^\star$ that

$$\int_{\mathbb{R}^n} \sigma_\delta(y)[\phi_t(y) - \phi_t^\star] \, \mathrm{d}y \leq \varepsilon, \quad \text{for all } \delta \in (0, \delta^\star]. \tag{51}$$

We verify this by splitting the integral into two parts. Since f is either L-Lipschitz or L-smooth, there is a quadratic polynomial $p\colon \mathbb{R} \to \mathbb{R}$ with nonnegative coefficients such that

$$\phi_t(y) \leq p(\|y - \xi^\star\|), \quad \text{for all } y \in \mathbb{R}^n, \tag{52}$$

and $p(0) = \phi_t^\star$. Thus, by the intermediate value theorem, we may fix $r \in (0,1)$ sufficiently small to ensure

$$p(r) - \phi_t^\star \leq \frac{\varepsilon}{2}. \tag{53}$$

This implies

$$\begin{aligned} \phi_t(y) - \phi_t^\star &\leq p(\|y - \xi^\star\|) - \phi_t^\star && \text{[54a]} \\ &\leq \frac{\varepsilon}{2}, \quad \text{for all } y \in \mathcal{B}(\xi^\star, r). && \text{[54b]} \end{aligned}$$

Thus, integrating over the ball $\mathcal{B}(\xi, r)$ reveals

$$\begin{aligned} A &\triangleq \int_{\mathcal{B}(\xi^\star,r)} \sigma_\delta(y)[\phi_t(y) - \phi_t^\star] \, \mathrm{d}y && \text{[55a]} \\ &\leq \int_{\mathcal{B}(\xi,r)} \sigma_\delta(y) \cdot \frac{\varepsilon}{2} \, \mathrm{d}y && \text{[55b]} \\ &\leq \frac{\varepsilon}{2} \cdot \int_{\mathbb{R}^n} \sigma_\delta(y) \, \mathrm{d}y && \text{[55c]} \\ &= \frac{\varepsilon}{2}, && \text{[55d]} \end{aligned}$$

where the second inequality follows from [26]. Next we integrate over the rest of $\mathbb{R}^n$. Define

$$\begin{aligned} B_\delta &\triangleq \int_{\mathbb{R}^n - \mathcal{B}(\xi^\star,r)} \sigma_\delta(y)[\phi_t(y) - \phi_t^\star] \, \mathrm{d}y && \text{[56a]} \\ &\leq \int_{\mathbb{R}^n - \mathcal{B}(\xi^\star,r)} \sigma_\delta(y) \cdot p(\|y - \xi^\star\|) \, \mathrm{d}y. && \text{[56b]} \end{aligned}$$

We may apply Lemma 1 to deduce there is $\omega > 0$ such that

$$B_\delta \leq \frac{\varepsilon}{2}, \quad \text{for all } \delta \in (0, \omega]. \tag{57}$$

Consequently, [55] and [57] together imply

$$\begin{aligned} \int_{\mathbb{R}^n} \sigma_\delta(y)[\phi_t(y) - \phi_t^\star] \, \mathrm{d}y &= A + B_\delta && \text{[58a]} \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} && \text{[58b]} \\ &\leq \varepsilon, \quad \text{for all } \delta \in (0, \omega]. && \text{[58c]} \end{aligned}$$

Hence [51] holds, taking $\delta^\star = \omega$, *i.e.* $\phi_t(\xi^\delta) \to \phi_t^\star$ as $\delta \to 0^+$.

Step 3 Let $\bar{\varepsilon} > 0$. It suffices to show there is $\bar{\delta} > 0$ such that

$$\|\xi^\delta - \xi^\star\| \leq \bar{\varepsilon}, \quad \text{for all } \delta \in (0, \bar{\delta}]. \tag{59}$$

Define

$$\mathcal{S} \triangleq \{z : \|z - \xi^\star\| \geq \bar{\varepsilon}\} \tag{60}$$

and note, by the strong convexity of ϕ_t (*e.g.* see [28]),

$$\phi_t(z) \geq \phi_t^\star + \frac{\theta \bar{\varepsilon}^2}{2}, \quad \text{for all } z \in \mathcal{S}. \tag{61}$$

By Step 2, there is $\mu > 0$ such that

$$\phi_t(\xi^\delta) \leq \phi_t^\star + \frac{\theta \bar{\varepsilon}^2}{4}, \quad \text{for all } \delta \in (0, \mu]. \tag{62}$$

Thus, $\xi^\delta \notin \mathcal{S}$, for all $\delta \in (0, \mu]$, *i.e.* (59) holds, taking $\bar{\delta} = \mu$. This completes the proof. □