A Hamilton-Jacobi-based Proximal Operator

Stanley Osher\textsuperscript{a,1}, Howard Heaton\textsuperscript{b,1}, and Samy Wu Fung\textsuperscript{a,2}

\textsuperscript{a}Dept. of Mathematics, University of California, Los Angeles; \textsuperscript{b}Typal Research, Typal LLC; \textsuperscript{c}Dept. of Applied Mathematics and Statistics, Colorado School of Mines

This manuscript was compiled on February 18, 2023

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 Monte Carlo 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 Operators and Moreau Envelopes

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

$$\text{prox}_t f (x) \triangleq \arg \min_{z \in \mathbb{R}^n} \left\{ f(z) + \frac{1}{2t} \| z - x \|^2 \right\} \quad [1]$$

and

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

where $\| \cdot \|$ denotes the $\ell_2$ norm. 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

$$\nabla u(x, t) = \frac{x - \text{prox}_t f(x)}{t}. \quad [3]$$

Rearranging reveals

$$\text{prox}_t f(x) = x - t \nabla u(x, t). \quad [4]$$

A key idea we use is to estimate the proximal for continuous $f$ by replacing $u$ with a smooth approximation $u^\delta \in C^\infty(\mathbb{R}^n)$, 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 time $T > 0$. For all times $t \in [0, T]$, the envelope $u$ is a viscosity solution (e.g. see (2, Chapter 3, Theorem 6)) to the HJ equation

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

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

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

where $\Delta u$ is the Laplacian of $u$. If $f$ is bounded and Lipschitz, Crandall and Lions (15) show $u^\delta$ approximates $u$, i.e. $u^\delta \to u$ uniformly as $\delta \to 0^+$.

Significance Statement

Many objectives do not admit explicit proximal formulas and cannot be estimated using exact gradients (e.g. when objectives are 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-mhj-prox

\textsuperscript{1}SO (Stanley Osher), HH (Howard Heaton), and SWF (Samy Wu Fung) contributed equally.

\textsuperscript{2}Correspondence should be addressed to SWF via email: swufung@mines.edu.
There are different ways to estimate the integral formula for $v^\delta$ with respect to $x$ reveals
\[
\nabla v^\delta(x,t) = \frac{1}{\delta t} \mathbb{E}_{\nu \sim N(x,\delta t)} [(x-y) \exp (-f(y)/\delta)] \quad \text{[13]}
\]
Plugging [12b] and [13] into [11] enables $\nabla u^\delta$ to be written as
\[
\nabla u^\delta(x,t) = \frac{1}{t} \left( x - \mathbb{E}_{\nu \sim N(x,\delta t)} [y \exp (-f(y)/\delta)] \right) \quad \text{[14]}
\]
The above relation was used in (19). Here we take a further step, combining [4] and [14] to get an HJ-based estimate
\[
\text{prox}_{\alpha f}(x) = x - t \nabla u^\delta(x,t) \quad \text{[15a]}
\]
\[
\approx x - t \nabla u^\delta(x,t) \quad \text{[15b]}
\]
\[
= \frac{\mathbb{E}_{\nu \sim N(x,\delta t)} [y \exp (-f(y)/\delta)]}{\mathbb{E}_{\nu \sim N(x,\delta t)} \exp (-f(y)/\delta)} \quad \text{[15c]}
\]
As shown below, Monte Carlo sampling enables efficient approximation of proximals in high dimensions (e.g., see Figure 2). Moreover, [15c] 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 or overflowing with limited numerical precision, due to either $\delta$ being small or $f(y)$ being large. Indeed, this makes the naïve implementation shown in Algorithm 1 numerically unstable. However, this may be remedied as the proximal formula may equivalently be re-scaled via
\[
\text{prox}_{\alpha f}(x) = \text{prox}_m \frac{\alpha f}{\epsilon}(x) \quad \text{[16a]}
\]
\[
\approx \frac{\mathbb{E}_{\nu \sim N(x,\delta t/\alpha)} [y \exp (-\alpha f(y)/\delta)]}{\mathbb{E}_{\nu \sim N(x,\delta t/\alpha)} \exp (-\alpha f(y)/\delta)} \quad \text{[16b]}
\]
where $t$ is replaced by $t/\alpha$ and $f$ by $\alpha f$ in [15c]. In this case, if $f/\delta$ becomes too large with respect to numerical precision limitations, it may be scaled down with a corresponding $\alpha$. We can check whether we obtain an underflow with $\exp(\alpha f(y)/\delta)$ and rescale $\alpha$ using a linesearch-like approach (e.g., see the support information where we add a single conditional statement to recursively halve $\alpha$ until $\exp(\alpha f(y)/\delta) > \epsilon$ for a tolerance $\epsilon$. Small $\alpha$ makes the variance large and more samples may be required to accurately estimate the expectations, i.e., a trade-off may be observed between numerical stability and accuracy of estimations. Another possible mitigation is to adaptively rescale $f$ based on the number of recursive steps taken in HJ-Prox. Note large $\delta$ can smooth approximations and mitigate the stochastic characteristics of HJ-Prox.

**Algorithm 1 HJ-Prox – Naïve Implementation**

1: HJ-Prox($x$, $t$; $f$, $\delta$, $N$, $\varepsilon$) :
2: for $i \in [N]$:
3: Sample $y^i \sim N(x,\delta t)$
4: $z_i \leftarrow f(y^i)$
5: prox $\leftarrow \text{softmax}(-z_i/\delta) \mathbb{E}^T [y^i \cdots y^N]$ 
6: return prox

This can greatly reduce sampling complexity (17, 18). Differentiating $v^\delta$ with respect to $x$ reveals

**Numerical Considerations**

A possible numerical challenge in our formulation is to address numerical instabilities arising from the exponential term underflowing or overflowing with limited numerical precision, due to either $\delta$ being small or $f(y)$ being large. Indeed, this makes the naïve implementation shown in Algorithm 1 numerically unstable. However, this may be remedied as the proximal formula may equivalently be re-scaled via

$$
\text{prox}_{\alpha f}(x) = \text{prox}_m \frac{\alpha f}{\epsilon}(x) \quad \text{[16a]}
$$

$$
\approx \frac{\mathbb{E}_{\nu \sim N(x,\delta t/\alpha)} [y \exp (-\alpha f(y)/\delta)]}{\mathbb{E}_{\nu \sim N(x,\delta t/\alpha)} \exp (-\alpha f(y)/\delta)} \quad \text{[16b]}
$$

where $t$ is replaced by $t/\alpha$ and $f$ by $\alpha f$ in [15c]. In this case, if $f/\delta$ becomes too large with respect to numerical precision limitations, it may be scaled down with a corresponding $\alpha$. We can check whether we obtain an underflow with $\exp(\alpha f(y)/\delta)$ and rescale $\alpha$ using a linesearch-like approach (e.g., see the support information where we add a single conditional statement to recursively halve $\alpha$ until $\exp(\alpha f(y)/\delta) > \epsilon$ for a tolerance $\epsilon$. Small $\alpha$ makes the variance large and more samples may be required to accurately estimate the expectations, i.e., a trade-off may be observed between numerical stability and accuracy of estimations. Another possible mitigation is to adaptively rescale $f$ based on the number of recursive steps taken in HJ-Prox. Note large $\delta$ can smooth approximations and mitigate the stochastic characteristics of HJ-Prox.
**Convergence Analysis**

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

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

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

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

\[
\lim_{\delta \to 0^+} \frac{E_{y \sim N(x, \delta t)}[\exp(-f(y)/\delta)]}{E_{y \sim N(x, \delta t)}[\exp(-f(y)/\delta)]} = \text{prox}_{tf}(x). \tag{17}
\]

A proof of Theorem 1 is in the supporting information (SI), and we note HJ-Prox may fail when \( f \) is discontinuous.

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

*Weakly convex functions are also referred to as semi-convex functions (20).

**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) (16, 19, 36–40).

The theoretical result 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 (16) and in the context of global optimization in (19).
Moreau Envelope for Nonconvex Functions

\[ f(x) = -|x|, \quad u(x) = -|x| - t/2 \]

\[ f(x) = -\frac{x^2}{2}, \quad u(x) = -\frac{x^2}{2(t+1)}, \quad t < 1 \]

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

Proximal Comparisons for Functions with Unknown Proximals

\[ f(x) = x^2 - \log(x) \]

Fig. 4. (a): Plots for function \( f \), exact Moreau envelope \( u \), and HJ-based Moreau envelope \( u^\delta \). (b): Plots for true proximal and approximate HJ-based proximal operators. (c): HJ-based Moreau envelopes \( u^\delta \) 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 Eqn. (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 \( C^\infty \) approximation of the underlying function \( f \) in the noisy case. Here, \( \delta = 0.1 \) for the noiseless case, and \( \delta_1 = 0.5 \) 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 \( u^\delta \) is a smooth (i.e., \( C^\infty(\mathbb{R}^n) \)) 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 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, 42)

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

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) (43) defines a sequence of solution estimates \( \{x^k\} \) for all \( k \in \mathbb{N} \) via

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

where the shrink operator defined element-wise by

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

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 \( \beta = 1/\|A\|_2 \). Our experiments show HJ-based ISTA can solve Lasso, up to an error tolerance.
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

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

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 [20]. Here, each oracle call returns

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

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 $\|Ax\|_1$. 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

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

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

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

$$^1$$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 $\|Ax\|_1$ 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 $x^*$, up to an error threshold, regardless of the choice of $\delta$ and number of samples $N$. Notice Figure 6a shows "small" values of $\delta$ 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 $u^k$ and averaging effect of Monte Carlo 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 viscosity solutions to Hamilton-Jacobi (HJ) equations, as given via the Hopf-Lax formula. Upon rewriting this approximation in terms of expectations, we use Monte Carlo 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 1952349.

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 Tvránič, 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 duality—rachfold 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 1-regularized problems. SIAM journal on imaging sciences 2, 323–343 (2009).
12. JJ Moreau, Decomposition 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. MG Grandall, PL Lions, Two approximations of solutions of Hamilton-Jacobi equations. Math. computation 43, 1–19 (1984).
16. P Daunhauri, A Oberman, S Osher, S Scasto, G Carlier, Deep relaxation: partial differential equations for optimizing deep neural networks. Res. Math. Sci. 5, 1–30 (2018).
17. 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).
18. ST Toldar, RE Kass, Importance sampling: a review. Wiley Interdiscip. Rev. Comput. Stat 2, 54–60 (2010).
19. H Heaton, SW Fung, S Osher, Global solutions to nonconvex problems by evolution of hamilton-jacobi pdes. arXiv preprint arXiv:2207.11014 (2022).
20. V Krytøt, J Zažíček, Differences of two semiconvex functions on the real line. Commentiones Math. Univ. Carol. 57, 21–37 (2016).
21. YM Ermolie, RB Wets, Numerical techniques for stochastic optimization. (Springer-Verlag), (1988).
22. D Kozak, S Becker, A Dosostan, L Tenorio, Stochastic subspace descent. arXiv preprint arXiv:1904.01145 (2019).
23. D Kozak, S Becker, A Dosostan, 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, D 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 (zero): Approximate sparse gradients and adaptive sampling. SIAM J. on Optim. 32, 687–714 (2022).
26. H Cai, Y Lou, DMcKenzie, 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 Slavín, 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 Menichelli, SM Wild, Derivative-free optimization methods. Acta Numer. 29, 287–404 (2019).
30. J Moré, S Wild, Benchmarking derivative-free optimization algorithms. SIAM J. on Optim. 20, 172–191 (2009).
31. JHM Shi, MQ Xuan, F Ozturak, J Nocedal, On the numerical performance of derivative-free optimization methods based on finite-difference approximations. arXiv preprint arXiv:2102.09762 (2021).
32. JHM Shi, Y Xie, MQ Xuan, J Nocedal, Adaptive finite-difference interval estimation for noisy derivative-free optimization. arXiv preprint arXiv:2110.13931 (2021).
33. B Kim, H Cai, D McKenzie, W Yin, Curvature-aware derivative-free optimization. arXiv preprint arXiv:2108.07890 (2021).
34. Lb Almeida, A learning rule for asynchronous perceptions with feedback in a combinatorial environment in Artificial neural networks: concept learning. pp. 102–111 (1996).
35. H Cai, D McKeen, 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 Barliar, 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 Diaz, D Drusvyatskiy, Escaping strict saddle points of the Moreau envelope in nonsmooth optimization. SIAM J. on Optim. 32, 1958–1983 (2022).
41. R Tishirnami, Regression shrinkage and selection via the lasso. J. Royal Stat. Soc. Ser. B (Methodological) 58, 267–288 (1996).
42. SS Chen, D Donoho, MA Saunders, Atomic decomposition by basis pursuit. SIAM review 43, 129–159 (2001).
43. Ar Bek, M Tsubole, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences 2, 183–202 (2009).
44. R Tibshirani, Regression shrinkage and selection via the lasso. J. Royal Stat. Soc. Ser. B (Methodological) 58, 267–288 (1996).
45. SS Chen, D Donoho, MA Saunders, Atomic decomposition by basis pursuit. SIAM review 43, 129–159 (2001).
Supporting Information: A Hamilton-Jacobi-based Proximal Operator

Stanley Osher\textsuperscript{a,1}, Howard Heaton\textsuperscript{b,1}, and Samy Wu Fung\textsuperscript{c,1,2}

\textsuperscript{a}Dept. of Mathematics, University of California, Los Angeles; \textsuperscript{b}Typal Research, Typal LLC; \textsuperscript{c}Dept. of Applied Mathematics and Statistics, Colorado School of Mines

HJ-Prox Implementation

Below we provide a more numerically stable HJ-Prox implementation that avoids underflow.

**Algorithm 1 HJ-Prox – IncludesUnderflow Check**

1. HJ-Prox\((x, t; f, \delta, N, \alpha, \varepsilon)\):
2. \hspace{1em} for \(i \in [N]\):
3. \hspace{2em} Sample \(y^i \sim N(x, \delta t/\alpha)\)
4. \hspace{2em} \(z_i \leftarrow f(y^i)\)
5. \hspace{2em} if \(\exp(-\alpha z_i/\delta) \leq \varepsilon\):
6. \hspace{3em} return HJ-Prox\((x, t; f, \delta, N, \alpha/2, \varepsilon)\)
7. \hspace{2em} prox \leftarrow \text{softmax}(\alpha z_i/\delta)^T [y^1 \cdots y^N]
8. return prox

Proofs

We fix the point \(x \in \mathbb{R}^n\) for the following calculations. For concise expression below, for \(t > 0\) and \(\delta > 0\), define

\[
\phi_t(z) \triangleq f(z) + \frac{1}{2t} \|z - x\|^2, \quad [1]
\]

\[
\phi^*_t \triangleq \inf \{\phi_t(y) : y \in \mathbb{R}^n\}, \quad \xi^* \triangleq \text{prox}_{\phi_t}(x) = \text{argmin } \phi_t \quad \text{(n.b. existence and uniqueness of the minimizer } \xi^* \text{ are shown below)}, \quad [2]
\]

and

\[
\sigma_\delta(z) \triangleq \frac{\exp(-\phi_t(z)/\delta)}{\|\exp(-\phi_t/\delta)\|_{L_1(\mathbb{R}^n)}} = e^{-\phi_t(z)/(\delta t)} \|e^{-\phi_t/\delta}\|_{L_1(\mathbb{R}^n)}. \quad [2]
\]

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

\[
\int_{\mathbb{R}^n} \sigma_\delta(y) \ dy = 1, \quad \sigma_\delta(y) \geq 0, \quad \text{for all } y \in \mathbb{R}^n, \quad [3]
\]

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

\[
\lim_{\delta \to 0^+} \int_{\mathbb{R}^n - B(\xi^*, r)} \sigma_\delta(y)p(\|y - \text{prox}_{\phi_t}(x)\|) \ dy = 0, \quad [4]
\]

where \(\text{prox}_{\phi_t}(x)\) is the unique minimizer of \(\phi_t\).

**Proof.** Since integration is linear and the limit of a sum is the sum of the limits, it suffices to verify [4] for any \(p(x) = x^k\) with \(k \geq 1\). First we show \(\sigma_\delta\) satisfies properties to be a probability density (Step 1). We also show various \(L^p\) norm limits hold for the denominator (Step 2) and numerator(Step 3) of the integrand in [4]. Combining these limits gives [4] (Step 4).

**Step 1** The numerator and denominator in the definition [2] for \(\sigma_\delta\) are nonnegative, making \(\sigma_\delta \geq 0\) everywhere. By the choice of \(t, \phi_t\) is \(\theta = 1/t - \rho\) strongly convex, and so it admits a unique minimizer \(\xi^* = \text{prox}_{\phi_t}(x)\) and satisfies

\[
\phi_t(y) \geq \phi^*_t + \langle 0, y - \xi^* \rangle + \frac{\theta}{2} \|y - \xi^*\|^2, \quad \text{for all } y \in \mathbb{R}^n. \quad [5]
\]

Consequently,

\[
0 < e^{-\phi_t(y)} \leq e^{-\phi^*_t + \frac{\theta}{2} \|y - \xi^*\|^2}, \quad \text{for all } y \in \mathbb{R}^n. \quad [6]
\]

Since the upper bound above is an exponential that decays quadratically (i.e. a Gaussian), the middle term in [6] is integrable over \(\mathbb{R}^n\), and so the denominator in the definition of \(\sigma_\delta\) is positive and finite. Thus, [3] readily follows as the integral of the numerator of \(\sigma_\delta\) equals the numerator of \(\sigma_\delta\).

**Step 2** A classic result in analysis (e.g. see (1, Exercise 3.4)) states \(L^p\) norms converge to the \(L^\infty\) norm as \(p \to \infty\), and so

\[
\lim_{\delta \to 0^+} \left\| e^{-\phi_t} \right\|_{L^\infty(\mathbb{R}^n)} = \left\| e^{-\phi^*_t} \right\|_{L^\infty(\mathbb{R}^n)} = e^{-\phi^*_t}, \quad [7]
\]

where, for all \(\delta > 0\), the \(L^\infty\) norm is finite by Step 1 and the final equality holds since \(\phi^*_t\) is the infimum of \(\phi_t\).

**Step 3** Integrating the numerator of [4] (i.e. not including division by the \(L^1\) norm in the definition of \(\sigma_\delta\)) for \(p(x) = x^k\) gives

\[
\int_{\mathbb{R}^n - B(\xi^*, r)} e^{-\phi_t(y)} \|y - \xi^*\|^k \ dy \quad [8a]
\]

\[
\leq \int_r^{\infty} e^{-\phi^*_t + \frac{\theta}{2} \|y - \xi^*\|^2} \cdot n|B(\xi^*, 1)| r^{k-1} \ dy \quad [8b]
\]

\[
= n|B(\xi^*, 1)| \int_r^{\infty} e^{-\phi^*_t + \frac{\theta}{2} \|y - \xi^*\|^2} \|y - \xi^*\|^k \ dy \quad [8c]
\]

where the first inequality holds by a change of variables to polar coordinates and using the strong convexity of \(\phi_t\) in [5], and the final line holds by properties of logarithms. Now define

\[
\varepsilon \triangleq \frac{\theta}{4(n + k - 1)} > 0, \quad [9]
\]
where the denominator is positive since \( n \geq 1 \) and \( p \) has positive degree (i.e. \( k \geq 1 \)). For all \( 0 < \delta < \varepsilon \), observe
\[
\tau > 1 \implies \tau^k < \tau^\varepsilon \quad \text{and} \quad \tau \leq 1 \implies \tau^k \leq 1^\varepsilon, \quad [10]
\]
i.e.
\[
\tau^\delta \leq \max(\tau, 1)^\varepsilon, \quad \text{for all} \; \delta \in (0, \varepsilon). \quad [11]
Whence, rewriting [8], we deduce, for all \( \delta \in (0, \varepsilon) \),
\[
\frac{1}{n|B(\xi^*, 1)|} \int_{\mathbb{R}^n - B(\xi^*, r)} e^{-\frac{\phi(y)}{\theta^2/2}} |y - \xi^*|^k \, dy \quad [12a]
\leq \int_0^\infty e^{-\frac{\theta^2}{2} \cdot (n + k - 1) \ln(\max(\gamma, 1))} \, d\tau. \quad [12b]
\]
Let \( q(\tau) \) be the numerator inside the exponential in the integrand of [12b]. Taking the limit yields
\[
\lim_{\delta \to 0^+} \left| \int e^{-\delta |y - \xi^*|^k} \, dy \right| = \left| \int e^{-\delta |y - \xi^*|^k} \, dy \right|_{L^\infty([r, \infty))} = 1. \quad [13]
\]
Let \( \tau^* \) be the minimizer of \( q \) over \([r, \infty)\). If \( \tau^* > 1 \), then the first order necessary condition and [9] together imply
\[
0 = \frac{\varepsilon(n + k - 1)}{\tau^*} \quad [14]
\]
and so
\[
\tau^* = \sqrt{\frac{\varepsilon(n + k - 1)}{\theta}} = \frac{1}{\sqrt{2}}. \quad [15]
\]
a contradiction (n.b. the second equality holds by choice of \( \varepsilon \) in [9]). Consequently, \( \tau^* \leq 1 \). Since \( q \) is quadratic in \( \tau \) and strictly increasing on \([r, 1)\), we deduce \( \tau^* = r \). Thus,
\[
|e^{-\delta |y - \xi^*|^k}|_{L^\infty([r, \infty))} = e^{-\frac{\theta^2}{2}}. \quad [16]
\]
Furthermore, note
\[
\lim_{\delta \to 0^+} \left| n|B(\xi^*, 1)|\right|^{\delta} = 1. \quad [17]
\]
Together [12], [16], and [17] imply
\[
\lim_{\delta \to 0^+} \left| \int e^{-\frac{\phi(y)}{\theta^2/2}} |y - \xi^*|^k \, dy \right|^{\delta} \leq e^{-\frac{\theta^2}{2}}. \quad [18]
\]
**Step 4** Define
\[
\gamma \triangleq \frac{e^{-\frac{\theta^2}{2}} - \frac{\theta^2}{2}}{e^{-\frac{\theta^2}{2}} - 2} \in (0, 1). \quad [19]
\]
By [7] and [18] and the definition of \( \sigma_\delta \),
\[
\lim_{\delta \to 0^+} \left| \int_{\mathbb{R}^n - B(\xi^*, r)} \sigma_\delta(y) |y - \xi^*|^k \, dy \right|^{\delta} \leq \gamma < 1. \quad [20]
\]
Consequently, there is \( \delta > 0 \) such that, for all \( \delta \in (0, \delta] \),
\[
\left| \int_{\mathbb{R}^n - B(\xi^*, r)} \sigma_\delta(y) |y - \xi^*|^k \, dy \right| \leq \frac{\gamma + 1}{2}, \quad [21]
\]
where we note \( (\gamma + 1)/2 \in (\gamma, 1) \), and so
\[
\lim_{\delta \to 0^+} \int_{\mathbb{R}^n} \sigma_\delta(y) |y - \xi^*|^k \, dy \leq \lim_{\delta \to 0^+} \left( \frac{\gamma + 1}{2} \right)^{1/\delta} = 0, \quad [22a]
\]
as desired.

**Lemma 2.** If the conditions of Theorem 1 hold, then there are constants \( a > 0 \) and \( b \geq 0 \) such that \( \phi_t \) has an upper bound of the form, for all \( y \in \mathbb{R}^n \),
\[
\phi_t(y) \leq a\|y - \text{prox}_f(x)\|^2 + b\|y - \text{prox}_f(x)\| + \phi_t^*, \quad [23]
\]
where \( \text{prox}_f(x) \) is the unique minimizer of \( \phi_t \).

Proof. For notational compactness, set \( \xi^* = \text{prox}_f(x) \), and note \( \xi^* \) exists and is unique by Lemma 1. We first verify the statement for \( L \)-Lipschitz \( f \) (Step 1) and then for when the gradient of \( f \) is \( L \)-Lipschitz (Step 2).

**Step 1** Suppose \( f \) is \( L \)-Lipschitz for some \( L > 0 \), i.e.
\[
\|f(y) - f(z)\| \leq L\|y - z\|, \quad \text{for all} \; y, z \in \mathbb{R}^n. \quad [24]
\]
Next note, for all \( y \in \mathbb{R}^n \),
\[
y - x - (\xi^* - x) \leq \frac{1}{2} \left\| \|y - x\|^2 - \|\xi^* - x\|^2 \right\| \quad [25a]
= \frac{1}{2} \left\| \|y - x\|^2 - (y - \xi^*) \right\| \quad [25b]
\leq \frac{1}{2} \left\| \|y - \xi^*\|^2 + 2\|y - \xi^*\||\|\xi^*\| + \|x\| \right\| \quad [25c]
= \frac{1}{2} \left\| \|y - \xi^*\|^2 + 2\|y - \xi^*\||\|\xi^*\| + \|x\| \right\|. \quad [25d]
\]
Consequently, [24] and [25] together imply, for all \( y \in \mathbb{R}^n \),
\[
\phi_t(y) - \phi_t^* \leq f(y) - f(\xi^*) + \frac{1}{2} \left\| \|y - x\|^2 - \|\xi^* - x\|^2 \right\| \quad [26a]
\leq L\|y - \xi^*\| \quad [26b]
\leq \frac{1}{2} \left\| \|y - \xi^*\|^2 + 2\|y - \xi^*\||\|\xi^*\| + \|x\| \right\|. \quad [26c]
\]
Thus, the upper bound in [23] holds with
\[
a = \frac{1}{2t} \quad \text{and} \quad b = \frac{\|\xi^*\| + \|x\|}{t} + L. \quad [27]
\]

**Step 2** Consider when \( f \) has an \( L \)-Lipschitz gradient for some \( L > 0 \). By (2, Lemma 5.7), for all \( y \in \mathbb{R}^n \),
\[
f(y) \leq f(\xi^*) + \langle \nabla f(\xi^*), y - \xi^* \rangle + \frac{L}{2} \|y - \xi^*\|^2 \quad [28a]
\leq f(\xi^*) + L\|y - \xi^*\| + \frac{L}{2} \|y - \xi^*\|^2. \quad [28b]
\]
Rearranging and again using [25] implies
\[
\phi_t(y) - \phi_t^* \leq L\|y - \xi^*\| + \frac{L}{2} \|y - \xi^*\|^2 \quad [29a]
\leq \frac{1}{2t} \left\| \|y - \xi^*\|^2 + 2\|y - \xi^*\||\|\xi^*\| + \|x\| \right\|. \quad [29b]
\]
Thus, the upper bound in [23] holds with
\[
a = \frac{1}{2t} \left( \frac{1}{L} + L \right) \quad \text{and} \quad b = L + \frac{\|\xi^*\| + \|x\|}{t}. \quad [30]
\]
This completes both cases of the proof.

\[\square\]
Below we restate and prove the main theorem, which is an extension of a lemma in Section 4.5.2 of (3).

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

\[
\lim_{\delta \to 0^+} \frac{E_{y \sim \mathcal{N}(x, \delta I)}}{E_{y \sim \mathcal{N}(x, \delta I)}} \left[ \exp \left( -f(y)/\delta \right) \right] = \text{prox}_{tf}(x). \quad [31]
\]

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

\[
\xi^\delta \triangleq \frac{E_{y \sim \mathcal{N}(x, \delta I)}}{E_{y \sim \mathcal{N}(x, \delta I)}} \left[ y \cdot \exp \left( -f(y)/\delta \right) \right], \quad \text{for all } \delta > 0, \quad [32]
\]

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

**Step 1** Note \( \xi^\delta \) can be rewritten via

\[
\xi^\delta = \left[ \int_{\mathbb{R}^n} e^{-\frac{\phi(y)}{\delta}} \, dy \right]^{-1} \int_{\mathbb{R}^n} y \cdot e^{-\frac{\phi(y)}{\delta}} \, dy. \quad [33]
\]

Using \( \sigma_{\delta} \), the estimate can be more concisely written via

\[
\xi^\delta = \int_{\mathbb{R}^n} \sigma_{\delta}(y) y \, dy = E_{y \sim \mathcal{N}(x, \delta I)}[y], \quad [34]
\]

where the expectation holds by utilizing the fact [3] shows \( \sigma_{\delta} \) defines a probability density. Thus, Jensen’s inequality may be applied to deduce

\[
\phi^*_t \leq \phi_t(\xi^\delta) = \phi_t(E_{y \sim \mathcal{N}(x, \delta I)}[y]) \leq E_{y \sim \mathcal{N}(x, \delta I)}[\phi_t(y)]. \quad [35]
\]

In integral form, we may subtract \( \phi^*_t \) to write

\[
0 \leq \phi_t(\xi^\delta) - \phi^*_t \leq \int_{\mathbb{R}^n} \sigma_{\delta}(y)[\phi_t(y) - \phi^*_t] \, dy. \quad [36]
\]

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

\[
|\phi_t(\xi^\delta) - \phi^*_t| \leq \varepsilon, \quad \text{for all } \delta \in (0, \delta^*]. \quad [37]
\]

By [36], the relation [37] holds if there is such a \( \delta^* \) that

\[
\int_{\mathbb{R}^n} \sigma_{\delta}(y)[\phi_t(y) - \phi^*_t] \, dy \leq \varepsilon, \quad \text{for all } \delta \in (0, \delta^*]. \quad [38]
\]

We verify this by splitting the integral into two parts. By Lemma 2, the fact \( f \) is either \( L \)-Lipschitz or \( L \)-smooth implies there is \( a > 0 \) and \( b \geq 0 \) such that, for all \( y \in \mathbb{R}^n \),

\[
\phi_t(y) - \phi^*_t \leq a\|y - \xi^*\|^2 + b\|y - \xi^*\|. \quad [39]
\]

Fix \( r \in (0, 1) \) sufficiently small to ensure

\[
r(ar + b) = ar^2 + br \leq \frac{\varepsilon}{2}. \quad [40]
\]

This implies

\[
\phi_t(y) - \phi^*_t \leq a\|y - \xi^*\|^2 + b\|y - \xi^*\| \leq \frac{\varepsilon}{2}, \quad \text{for all } y \in \mathcal{B}(\xi^*, r). \quad [41a]
\]

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

\[
A \triangleq \int_{\mathcal{B}(\xi^*, r)} \sigma_{\delta}(y)[\phi_t(y) - \phi^*_t] \, dy \leq \frac{\varepsilon}{2} \quad [41b]
\]

By Lemma 1, there is \( \omega > 0 \) such that

\[
B_{\delta} \leq \frac{\varepsilon}{2}, \quad \text{for all } \delta \in (0, \omega). \quad [44]
\]

Consequently, [42] and [44] together imply

\[
\int_{\mathbb{R}^n} \sigma_{\delta}(y)[\phi_t(y) - \phi^*_t] \, dy = A + B_{\delta} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \quad [45a]
\]

\[
\leq \varepsilon, \quad \text{for all } \delta \in (0, \omega]. \quad [45b]
\]

Hence [38] holds, taking \( \delta^* = \omega \). That is, we obtain the convergence \( \phi_t(\xi^\delta) \to \phi^*_t \) as \( \delta \to 0^+ \).

**Step 3** Let \( \overline{\delta} > 0 \). It suffices to show there is \( \overline{\delta} > 0 \) such that

\[
\|\xi^\delta - \xi^*\| \leq \overline{\delta}, \quad \text{for all } \delta \in (0, \overline{\delta}). \quad [46]
\]

Define

\[
S \triangleq \{ z : \|z - \xi^*\| \geq \overline{\delta} \}, \quad [47]
\]

and note, by the strong convexity of \( \phi_t \) (e.g. see [5]),

\[
\phi_t(z) \geq \phi_t^* + \frac{\theta r^2}{2}, \quad \text{for all } z \in S. \quad [48]
\]

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

\[
\phi_t(\xi^\delta) \leq \phi_t^* + \frac{\theta r^2}{4}, \quad \text{for all } \delta \in (0, \mu]. \quad [49]
\]

Thus, \( \xi^\delta \notin S \), for all \( \delta \in (0, \mu] \), i.e. (46) holds, taking \( \overline{\delta} = \mu \). This completes the proof.

---

1. W. Rudin, Real and Complex Analysis. (McGraw-Hill), (1966).
2. A. Beck, First-order methods in optimization. (SIAM), (2017).
3. L.C. Evans, Partial Differential Equations. Graduate Stud. Math. 19 (2010).