A Monte Carlo Approach for Nonsmooth Convex Optimization via Proximal Splitting Algorithms

Nicholas Di ND56@RICE.EDU
Rice University

Eric C. Chi ECHI@UMN.EDU
University of Minnesota

Samy Wu Fung SWUFUNG@MINES.EDU
Colorado School of Mines

Abstract

Operator splitting algorithms are a cornerstone of modern first-order optimization, relying critically on proximal operators as their fundamental building blocks. However, explicit formulas for proximal operators are available only for limited classes of functions, restricting the applicability of these methods. Recent work introduced HJ-Prox [18], a zeroth-order Monte Carlo approximation of the proximal operator derived from Hamilton–Jacobi PDEs, which circumvents the need for closed-form solutions. In this work, we extend the scope of HJ-Prox by establishing that it can be seamlessly incorporated into operator splitting schemes while preserving convergence guarantees. In particular, we show that replacing exact proximal steps with HJ-Prox approximations in algorithms such as proximal gradient descent, Douglas–Rachford splitting, Davis–Yin splitting, and the primal–dual hybrid gradient method still ensures convergence under mild conditions.

Keywords: Proximal, Operator Splitting, Derivative-Free, Zeroth-Order, Optimization, Monte Carlo, Hamilton-Jacobi

1. Introduction

In modern machine learning and optimization, splitting algorithms play an important role in solving complex problems, particularly those with nonsmooth composite objective functions [19]. Splitting algorithms face difficulty when a step involving the proximal operator lacks a closed-form solution, calling for computationally expensive and complex inner iterations to solve sub-optimization problems [22, 23]. To address this challenge, we build on recent work of Osher, Heaton, and Wu Fung, who showed that Hamilton–Jacobi (HJ) equations can be used to approximate proximal operators via a Monte Carlo scheme, termed HJ-Prox. [18]. In this work, we propose a new framework for splitting algorithms that replace the exact proximal operator with the HJ-Prox approximation. Our primary contribution is a theoretical and empirical demonstration that this new general framework maintains convergence near the true solution, reducing the need for proximal calculus and introducing a more universal and readily applicable approach to splitting algorithms. For this workshop paper, we focus on proximal gradient descent (PGD), Douglas Rachford Splitting (DRS), Davis-Yin Splitting (DYS), and the primal-dual hybrid gradient algorithm (PDHG) [21].

2. Background

Splitting algorithms are designed to solve composite convex optimization problems of the form

$$\min_x f(x) + g(x), \tag{1}$$

f and g are proper, lower-semicontinuous (LSC) and convex. Their efficiency, however, depends critically on the availability of closed-form proximal operators. When these operators are unavailable, the proximal step must be approximated through iterative subroutines, creating a substantial computational bottleneck. To address this challenge, several lines of research have emerged. One approach focuses on improving efficiency through randomization within the algorithmic structure. These methods reduce computational cost by sampling blocks of variables, probabilistically skipping the proximal step, or solving suboptimization problems incompletely with controlled error [1, 2, 6, 17]. Alternatively, other approaches reformulate the problem by focusing on dual formulations [13, 22].

While these techniques improve efficiency, they share common limitations. They require extensive derivations and complex analysis to handle the proximal operator, and proximal operators remain problem-dependent, typically requiring tailored solution strategies for each specific function class. This creates a critical research gap: the need for a generalizable method that can approximate the proximal operator without derivative information, making it suitable for zeroth-order optimization problems where only function evaluations are available.

2.1. Hamilton-Jacobi-based Proximal (HJ-Prox)

A promising solution to this challenge has emerged from recent work that approximates the proximal operator using a Monte-Carlo approach inspired by Hamilton-Jacobi (HJ) PDEs. Specifically, Osher, Heaton, and Wu Fung [18] showed that

$$\begin{aligned}
\operatorname{prox}_{tf}(x) &= \lim_{\delta \to 0^+} \frac{\mathbb{E}_{y \sim \mathcal{N}(x, \delta t I)} \left[y \cdot \exp(-f(y)/\delta) \right]}{\mathbb{E}_{y \sim \mathcal{N}(x, \delta t I)} \left[\exp(-f(y)/\delta) \right]} && (2) \\
&\approx \frac{\mathbb{E}_{y \sim \mathcal{N}(x, \delta t I)} \left[y \cdot \exp(-f(y)/\delta) \right]}{\mathbb{E}_{y \sim \mathcal{N}(x, \delta t I)} \left[\exp(-f(y)/\delta) \right]} \quad \text{for some } \delta > 0 && (3) \\
&= \operatorname{prox}^{\delta}_{tf}(x) && (4)
\end{aligned}$$

where $\mathcal{N}(x, \delta t I)$ represents the normal distribution with mean x and covariance matrix $\delta t I$, $t > 0$, and f is assumed to be weakly-convex [21].

The HJ-Prox, denoted by $\operatorname{prox}^{\delta}_{tf}$ in (4), fixes a small value of $\delta > 0$ to approximate the limiting expression above, enabling a Monte Carlo approximation of the proximal operator in a zeroth-order manner [11, 16, 18, 24].

This approach is particularly attractive because it requires only function evaluations, avoiding the need for derivatives or closed-form solutions. Subsequent research has explored HJ-Prox applications, primarily in global optimization via adaptive proximal point algorithms [11, 25, 26]. However, these applications have remained narrow in scope, focusing on specific algorithmic contexts rather than establishing a general framework. Our work expands upon the theory of HJ-Prox by creating a comprehensive framework that can be applied to the entire family of splitting algorithms for convex optimization, including proximal gradient descent (PGD) [20, 21], Douglas

Rachford Splitting (DRS) [8, 12], Davis-Yin Splitting (DYS) [7], and primal-dual hybrid gradient (PDHG) [3]. To our knowledge, the direct approximation of the proximal operator via HJ equations for use in general splitting methods has not been previously explored.

3. HJ-Prox-based Operator Splitting

We now show how HJ-Prox can be incorporated into splitting algorithms such as PGD, DRS, DYS, and PDHG. The key idea is simple: by replacing exact proximal steps with their HJ-Prox approximations, we retain convergence guarantees while eliminating the need for closed-form proximal formulas or costly inner optimization loops. For readability, all proofs are deferred to the Appendix.

Our analysis builds on a classical result concerning perturbed fixed-point iterations. In particular, Combettes [5, Thm. 5.2] established convergence of Krasnosel'skiĭ–Mann (KM) iterations subject to summable errors:

Theorem 1 (Convergence of Perturbed Krasnosel'skii-Mann Iterates) *Let $\{x_k\}_{k\geq 0}$ be an arbitrary sequence generated by*

$$x_{k+1} \quad = \quad x_k + \lambda_k \left(Tx_k + \epsilon_k - x_k\right), \tag{5}$$

where $T\colon \mathbb{R}^n \to \mathbb{R}^n$ is an operator that has at least one fixed point. If $\{\|\epsilon_k\|\}_{k\geq 0} \in \ell^1$ (that is, ϵ_k is summable), $T - I$ is demiclosed at 0, and $\{\lambda_k\}_k \geq 0$ lies in $[\gamma, 2-\gamma]$ for some $\gamma \in (0,1)$, then $\{x_k\}_{k\geq 0}$ converges to a fixed point of T.

Thus, to establish convergence of HJ-Prox–based splitting, it suffices to bound the HJ approximation error. The following result, originally proved in [18, 25], provides the required bound.

Theorem 2 (Error Bound on HJ-Prox) *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be LSC. Then the Hamilton-Jacobi approximation incurs errors that are uniformly bounded.*

$$\sup_x \left\| \operatorname{prox}_{tf}^{\delta}(x) - \operatorname{prox}_{tf}(x) \right\| \quad \leq \quad \sqrt{nt\delta}. \tag{6}$$

This uniform error bound guides the choice of δ in each iteration of our splitting algorithms. In particular, by selecting δ_k so that the resulting error sequence is summable, Theorem 1 ensures convergence of the HJ-Prox–based methods.

We rely on Theorem 1 to prove the convergence of the four fixed-point methods that use HJ-Prox. For simplicity, take $\lambda_k = 1$ for all k. Recall that $T - I$ is demiclosed at 0 if T is nonexpansive. In addition, recall that T is nonexpansive if T is averaged. Associated with each algorithm of interest is an algorithm map T that takes the current iterate x_k to the next iterate x_{k+1}. Consequently, to invoke Theorem 1, we verify that T is averaged and check the summability of the error introduced into T by the HJ-prox approximation.

Theorem 3 (HJ-Prox PGD) *Let f, g be proper, LSC, and convex, with f additionally L-smooth. Consider the HJ-Prox-based PGD iteration given by*

$$x_{k+1} \quad = \quad \operatorname{prox}_{tg}^{\delta_k}\left(x_k - t\nabla f(x_k)\right), \quad k = 1, \ldots, \tag{7}$$

with step size $0 < t < 2/L$ and $\left\{\sqrt{\delta_k}\right\}_{k\geq 1}$ a summable sequence. Then x_k converges to a minimizer of $f + g$.

Theorem 4 (HJ-Prox DRS) *Let f, g be proper, convex, and LSC. Consider the HJ-Prox–based DRS iteration given by*

$$\begin{aligned} x_{k+1/2} &= \operatorname{prox}_{tf}^{\delta_k}(z_k), \\ x_{k+1} &= \operatorname{prox}_{tg}^{\delta_k}(2x_{k+1/2} - z_k), \\ z_{k+1} &= z_k + x_{k+1} - x_{k+1/2}, \end{aligned} \tag{8}$$

with $\left\{\sqrt{\delta_k}\right\}_{k\geq 1}$ a summable sequence. Then x_k converges to a minimizer of $f+g$.

Theorem 5 (HJ-Prox DYS) *For DYS, consider $f+g+h$. Let f, g, h be proper, LSC, and convex, with h additionally L-smooth. Consider the HJ-Prox–based DYS algorithm given by*

$$\begin{aligned} y_{k+1} &= \operatorname{prox}_{tf}^{\delta_k}(x_k), \\ z_{k+1} &= \operatorname{prox}_{tg}^{\delta_k}\big(2y_{k+1} - x_k - t\nabla h(y_{k+1})\big) \\ x_{k+1} &= x_k + z_{k+1} - y_{k+1} \end{aligned} \tag{9}$$

with $\{\sqrt{\delta_k}\}_{k\geq 1}$ a summable sequence, and $0 < t < 2/L$. Then x_k converges to a minimizer of $f+g+h$.

Theorem 6 (HJ-Prox PDHG) *Let f, g be proper, convex, and LSC. Consider the HJ-Prox–based PDHG algorithm given by*

$$\begin{aligned} y_{k+1} &= \operatorname{prox}_{\sigma g^*}^{\delta_k}(y_k + \sigma A x_k), \\ x_{k+1} &= \operatorname{prox}_{\tau f}^{\delta_k}(x_k - \tau A^\top y_{k+1}), \end{aligned} \tag{10}$$

with parameters $\tau, \sigma > 0$ satisfying $\tau\sigma\|A\|^2 < 1$ and $\{\sqrt{\delta_k}\}_{k\geq 1}$ a summable sequence. Where g^* denotes the Fenchel conjugate of g. Then x^k converges to a minimizer of $f(x) + g(Ax)$.

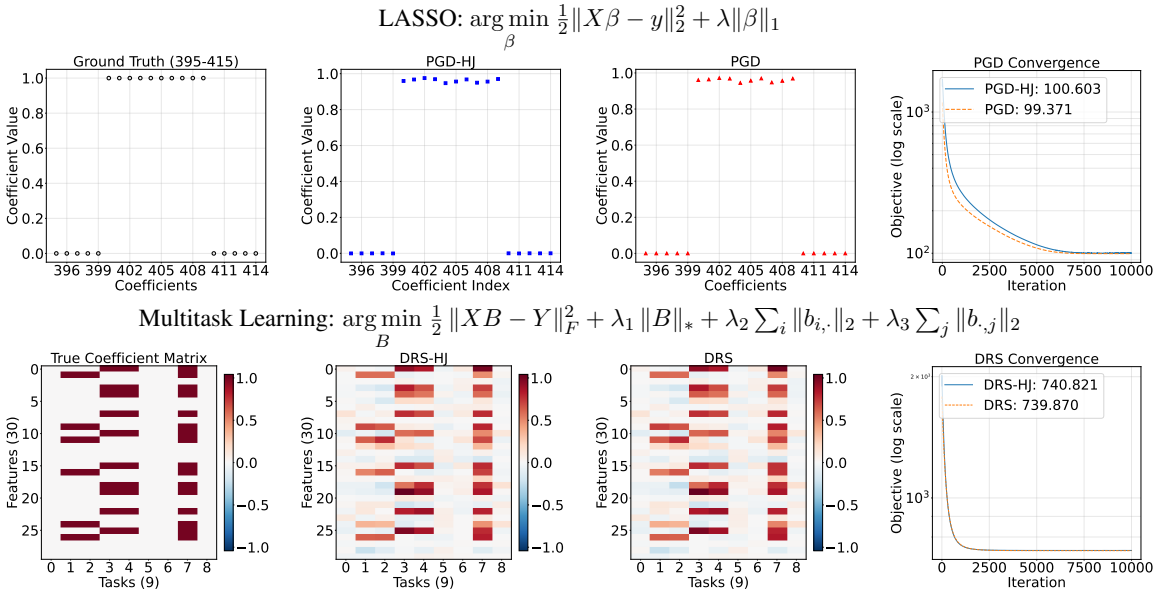

Figure 1: LASSO and Multitask Learning Results

4. Experiments

We conduct five experiments to assess the effectiveness of HJ-Prox integrated with proximal splitting methods. First, we use PGD to solve the LASSO, an ℓ_1-regularized least-squares model for sparse feature selection displayed in figure 1. We then apply DRS to multitask learning for structured sparse matrix recovery and to a third-order fused-LASSO formulation on Doppler data that smooths the underlying signal by penalizing third-order finite differences. Results are shown in Figures 1 and 2. Next, we solve the sparse group LASSO with DYS, which induces sparsity both across groups and within groups displayed in Figure 2. Finally, in Figure 3, we use the PDHG method for total-variation (TV) image denoising, preserving sharp edges by penalizing the ℓ_1 norm of the image gradients on a noisy, blurred sample image. We use identical problem parameters for both the HJ-Prox and analytical methods. For each experiment, the HJ-Prox temperature parameter δ_k is scheduled to satisfy conditions for convergence established in Theorem 1. Each experiment is designed to visually compare respective recovered signals with the ground truth. We also report the convergence and last iteration of the objective function values in the legend for both approaches. Further experimental details are in the Appendix F.

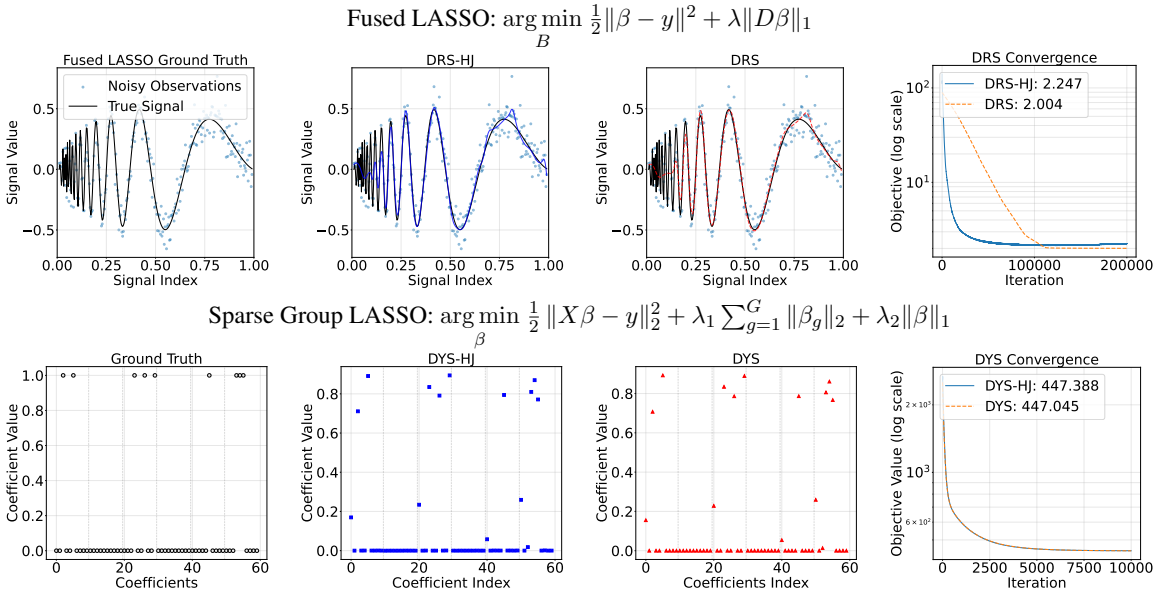

Figure 2: Fused LASSO and Sparse Group LASSO Results

4.1. Results

Across all experiments, HJ-Prox tracks the analytical baselines closely and converges to visually indistinguishable solutions. For LASSO and sparse group LASSO, the method performs effective variable selection, shrinking true zero coefficients toward zero as seen in Figures 1 and 2. In multitask learning and sparse group LASSO, the HJ-Prox iterates closely match the analytical updates. For fused LASSO in Figure 2, HJ-Prox exhibits faster initial convergence but settles farther away than the analytical solver, likely due to differences in formulation (primal vs dual) and the aggressive delta schedule for this particularly challenging proximal operator. We note that convergence speed

is problem dependent, as seen in Figures 2 and 3, TV and fused LASSO typically require more iterations due to additional subproblems and higher per iteration cost. We note that our goal is not to outperform specialized solvers but to demonstrate that a universal zeroth-order, sampling based proximal approximation integrated with standard splitting algorithms recovers the same solutions with analytical counterparts.

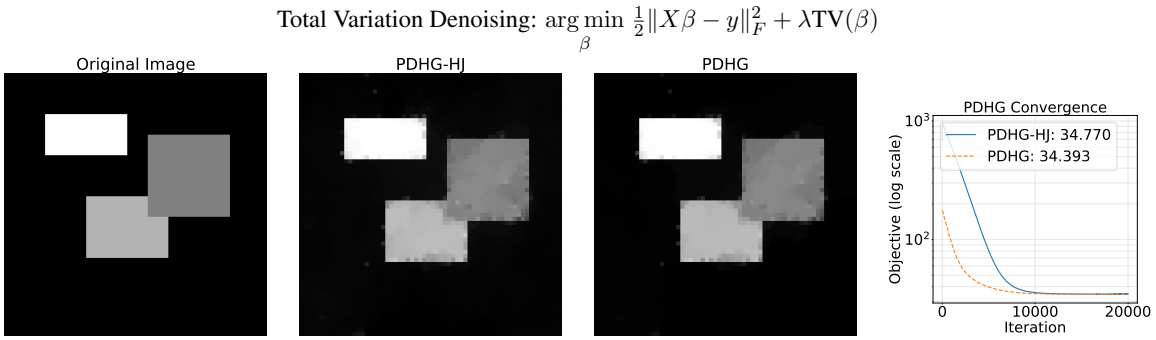

Figure 3: Total Variation Results

5. Limitations and Future Work

Several limitations and promising directions emerge from our analysis. First, our theoretical framework does not account for Monte Carlo sampling errors and assumes exact integral evaluation. To provide more realistic performance guarantees in practical applications, we plan to extend our analysis to incorporate finite-sample errors. Second, convergence can be slow when using set δ_k schedules and fixed step and sample sizes. Observed in our fused LASSO experiments, adaptive algorithm parameters are often necessary for efficient convergence. We suspect that jointly adapting both the sample size N and δ throughout the iterative process could be more effective than our current approach of using predetermined schedules as is done for proximal point in [11, 25]. We aim to develop adaptive splitting algorithms that dynamically adjust parameters based on problem-specific behavior. Future work also includes integrating HJ-Prox-based algorithms within a Learning-to-Optimize framework [4, 10, 14, 15] to enable automatic tuning of N and δ.

6. Conclusion

Our work demonstrates that HJ-Prox can be successfully integrated into operator splitting frameworks while maintaining theoretical convergence guarantees, providing a generalizable method for solving composite convex optimization problems. By replacing exact proximal operators with a zeroth-order Monte Carlo approximation, we have established that algorithms such as PGD, DRS, DYS, and the PDHG method retain their convergence properties under mild conditions. This framework offers practitioners a universal approach to solving complex non-smooth optimization problems, reducing reliance on expensive and complex proximal computations. Our code for experimentation will be available upon publication.

References

[1] Silvia Bonettini, Marco Prato, and Simone Rebegoldi. Convergence of inexact forward–backward algorithms using the forward–backward envelope. *Optimization Online (preprint)*, 2020. URL https://optimization-online.org/2020/02/7644/.

[2] Luis M. Briceño-Arias, Giovanni Chierchia, Emilie Chouzenoux, and Jean-Christophe Pesquet. A random block-coordinate Douglas–Rachford splitting method with low computational complexity for binary logistic regression. *Computational Optimization and Applications*, 72(3):707–726, 2019. doi: 10.1007/s10589-019-00060-6. URL https://link.springer.com/article/10.1007/s10589-019-00060-6.

[3] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40 (1):120–145, 2011. doi: 10.1007/s10851-010-0251-1. URL https://hal.science/hal-00490826/file/pd_alg_final.pdf.

[4] Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *Journal of Machine Learning Research*, 23(189):1–59, 2022.

[5] Patrick L. Combettes. Quasi-Fejérian analysis of some optimization algorithms. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, volume 8 of *Studies in Computational Mathematics*, pages 115–152. Elsevier, Amsterdam, The Netherlands, 2001.

[6] Laurent Condat and Peter Richtárik. Randprox: Primal–dual optimization algorithms with randomized proximal updates. *Optimization Online (preprint)*, 2022. URL https://arxiv.org/abs/2207.12891.

[7] Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4):829–858, 2017. doi: 10.1007/s11228-017-0421-z. URL https://doi.org/10.1007/s11228-017-0421-z.

[8] Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical programming*, 55(1):293–318, 1992.

[9] Olivier Fercoq. Quadratic error bound of the smoothed gap and the restarted averaged primal-dual hybrid gradient. *arXiv preprint arXiv:2206.03041*, 2022. doi: 10.48550/arXiv.2206.03041. URL https://arxiv.org/pdf/2206.03041.

[10] Howard Heaton and Samy Wu Fung. Explainable AI via learning to optimize. *Scientific Reports*, 13(1):10103, 2023.

[11] Howard Heaton, Samy Wu Fung, and Stanley Osher. Global solutions to nonconvex problems by evolution of Hamilton-Jacobi PDEs. *Communications on Applied Mathematics and Computation*, 6(2):790–810, 2024.

[12] Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[13] Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *arXiv preprint arXiv:1111.5479*, 2012. doi: 10.48550/arXiv.1111.5479. URL https://arxiv.org/abs/1111.5479.

[14] Daniel McKenzie, Howard Heaton, and Samy Wu Fung. Differentiating through integer linear programs with quadratic regularization and davis-yin splitting. *Transactions on Machine Learning Research*, 2024.

[15] Daniel Mckenzie, Howard Heaton, Qiuwei Li, Samy Wu Fung, Stanley Osher, and Wotao Yin. Three-operator splitting for learning to predict equilibria in convex games. *SIAM Journal on Mathematics of Data Science*, 6(3):627–648, 2024.

[16] Tingwei Meng, Siting Liu, Samy Wu Fung, and Stanley Osher. Recent advances in numerical solutions for Hamilton-Jacobi PDEs. *arXiv preprint arXiv:2502.20833*, 2025.

[17] Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtarik. Proxskip: Yes! Local gradient steps provably lead to communication acceleration! Finally! In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15750–15769. PMLR, July 17–23 2022. URL https://proceedings.mlr.press/v162/mishchenko22b.html.

[18] Stanley Osher, Howard Heaton, and Samy Wu Fung. A Hamilton–Jacobi-based proximal operator. *Proceedings of the National Academy of Sciences of the United States of America*, 120(14):e2220469120, 2023. doi: 10.1073/pnas.2220469120. URL https://www.pnas.org/doi/10.1073/pnas.2220469120.

[19] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014. doi: 10.1561/2400000003.

[20] R. Tyrell Rockafellar. Convex analysis. *Princeton Mathematical Series*, 28, 1970.

[21] Ernest K. Ryu and Wotao Yin. *Large-scale convex optimization: algorithms & analyses via monotone operators*. Cambridge University Press, 2022.

[22] Ryan J. Tibshirani. Dykstra's algorithm, ADMM, and coordinate descent: Connections, insights, and extensions. In *Advances in Neural Information Processing Systems*, NeurIPS, 2017. URL https://www.stat.berkeley.edu/~ryantibs/papers/dykcd.pdf. Submitted at NeurIPS 2017.

[23] Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011. doi: 10.1214/11-AOS878. URL https://projecteuclid.org/journals/annals-of-statistics/volume-39/issue-3/The-solution-path-of-the-generalized-lasso/10.1214/11-AOS878.full.

[24] Ryan J. Tibshirani, Samy Wu Fung, Howard Heaton, and Stanley Osher. Laplace meets Moreau: Smooth approximation to infimal convolutions using Laplace's method. *Journal of Machine Learning Research*, 26(72):1–36, 2025.

[25] Minxin Zhang, Fuqun Han, Yat Tin Chow, Stanley Osher, and Hayden Schaeffer. Inexact proximal point algorithms for zeroth-order global optimization. *arXiv preprint arXiv:2412.11485*, 2024. doi: 10.48550/arXiv.2412.11485. URL https://arxiv.org/abs/2412.11485.

[26] Zhiwei Zhang, Samy Wu Fung, Anastasios Kyrillidis, Stanley Osher, and Moshe Y Vardi. Thinking out of the box: Hybrid sat solving by unconstrained continuous optimization. *arXiv preprint arXiv:2506.00674*, 2025.

Appendix A. Proof of HJ-Prox Error Bound

For completeness and ease of presentation, we restate the theorem.

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be LSC. Then the Hamilton-Jacobi approximation incurs errors that are uniformly bounded.

$$\sup_x \left\| \text{prox}_{tf}^{\delta}(x) - \text{prox}_{tf}(x) \right\| \quad \leq \quad \sqrt{nt\delta}. \tag{11}$$

Proof Fix the parameters t and δ. For notational convenience, denote $\text{prox}_{tf}(x)$ by $z^\star(x)$ and $\text{prox}_{tf}^{\delta}(x)$ by $z_\delta(x)$. Let

$$\phi_x(z) \quad = \quad f(z) + \frac{1}{2t}\|x - z\|^2. \tag{12}$$

Making the change of variable $w = z - z^\star(x)$ in (4) enables us to expresses the approximation error as

$$z_\delta(x) - z^\star(x) \quad = \quad \frac{\int w \exp\left(-\frac{\phi_x(z^\star(x)+w)-\phi_x(z^\star(x))}{\delta}\right) dw}{\int \exp\left(-\frac{\phi_x(z^\star(x)+w)-\phi_x(z^\star(x))}{\delta}\right) dw}. \tag{13}$$

Let

$$Z_\delta \quad = \quad \int \exp\left(-\frac{\phi_x(z^\star(x)+w) - \phi_x(z^\star(x))}{\delta}\right) dw \tag{14}$$

and

$$g(w) \quad = \quad \phi_x(z^\star(x)+w) - \phi_x(z^\star(x)). \tag{15}$$

Then

$$\rho_\delta(w) \quad = \quad \frac{e^{-\frac{g(w)}{\delta}}}{Z_\delta} \tag{16}$$

defines a proper density.

Equations (13) and (16) together imply that the approximation error can be written as the expected value of a continuous random variable W whose probability law has the density ρ_δ.

$$z_\delta(x) - z^\star(x) \quad = \quad \int w\, \rho_\delta(w) dw \quad = \quad \mathbb{E}_{\rho_\delta}(W). \tag{17}$$

Taking the norm of both sides of (17) leads to a bound on the norm of the approximation error.

$$\|z_\delta(x) - z^\star(x)\| \quad = \quad \|\mathbb{E}_{\rho_\delta}(W)\| \quad \leq \quad \mathbb{E}_{\rho_\delta}(\|W\|) \quad \leq \quad \sqrt{\mathbb{E}_{\rho_\delta}\left(\|W\|^2\right)}. \tag{18}$$

The first inequality is due to Jensen's inequality since norms are convex. The second is due to the Cauchy-Schwarz inequality.

Our goal is to show that

$$\sqrt{\mathbb{E}_{\rho_\delta}\left(\|W\|^2\right)} \quad \leq \quad \sqrt{nt\delta}, \tag{19}$$

since inequalities (18) and (19) together imply that

$$\|z_\delta(x) - z^\star(x)\| \quad \leq \quad \sqrt{nt\delta}. \tag{20}$$

We prove (19) in two steps. We first show that

$$\mathbb{E}_{\rho_\delta}\left(\|W\|^2\right) \quad \leq \quad t\mathbb{E}_{\rho_\delta}\left(\langle W, \nabla g(W)\rangle\right), \tag{21}$$

where g is the convex function defined in (15). We then show that

$$\mathbb{E}_{\rho_\delta}\left(\langle W, \nabla g(W)\rangle\right) \quad = \quad n\delta. \tag{22}$$

Before proceeding to prove these steps, we address our abuse of notation in (21) and (22). Although ∇g may not exist everywhere, it exists almost everywhere. Recall that g is locally Lipschitz because it is convex. Furthermore, any locally Lipschitz function is differentiable almost everywhere by Rademacher's theorem. Hence ∇g exists almost everywhere. Consequently, the expectation $\mathbb{E}_{\rho_\delta}\langle W, \nabla g(W)\rangle$ is well defined.

To show (21), first note that by Fermat's rule, $0 \in \partial\phi_x(z^\star(x))$ where $\partial\phi(z)$ denotes the subdifferential of ϕ at z. Consequently, for any $z \in \mathbb{R}^n$

$$\begin{aligned} \phi_x(z) \quad &\geq \quad \phi_x(z^\star(x)) + \langle 0, z - z^\star(x)\rangle + \frac{1}{2t}\|z - z^\star(x)\|^2 \\ &= \quad \phi_x(z^\star(x)) + \frac{1}{2t}\|z - z^\star(x)\|^2, \end{aligned} \tag{23}$$

since $\phi_x(z)$ is $\frac{1}{t}$-strongly convex. Plugging $z = z^\star(x) + w$ into inequality (23) implies that

$$g(w) \quad \geq \quad \frac{1}{2t}\|w\|^2. \tag{24}$$

We also know that since $\phi_x(z)$ is $\frac{1}{t}$-strongly convex, so is $g(z)$, since g is a translation of ϕ plus a constant shift. For $\frac{1}{t}$ strongly convex g with $\nabla g(w)$ existing at point w, and setting $w' = 0$, the strong convexity inequality gives

$$g(0) \quad \geq \quad g(w) + \langle \nabla g(w), 0 - w\rangle + \frac{1}{2t}\|0 - w\|^2. \tag{25}$$

Therefore,

$$\langle \nabla g(w), w\rangle \quad \geq \quad g(w) + \frac{1}{2t}\|w\|^2. \tag{26}$$

Using (24) and (26),

$$\langle \nabla g(w), w\rangle \geq \frac{1}{t}\|w\|^2. \tag{27}$$

which implies that

$$\frac{1}{t}\int \|w\|^2 \rho_\delta(w)dw \quad \leq \quad \int \langle w, \nabla g(w)\rangle \rho_\delta(w)dw. \tag{28}$$

To show (22), consider w where ∇g exists and let $h(w) = \exp(-g(w)/\delta)$. By the chain rule

$$\frac{\partial}{\partial w_j} g(w)h(w) \quad = \quad -\delta \frac{\partial}{\partial w_j} h(w). \tag{29}$$

Integrating both sides of (29) over $\mathbb{R}^d$ gives

$$\begin{aligned} \int_{\mathbb{R}^n} w_j \frac{\partial}{\partial w_j} g(w)h(w)dw &= -\delta \int_{\mathbb{R}^n} w_j \frac{\partial}{\partial w_j} h(w)dw \\ &= -\delta \int_{\mathbb{R}^{n-1}} \left[\int_{-\infty}^{\infty} w_j \frac{\partial}{\partial w_j} h(w)dw_j \right] dw_{-j}, \end{aligned} \tag{30}$$

where w_{-j} is the subvector of w containing all but its jth element.

Applying integration by parts on the right hand side of (30) gives

$$\int_{-\infty}^{\infty} w_j \frac{\partial}{\partial_{w_j}} h(w)dw_j \quad = \quad w_j h(w)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} h(w)dw_j. \tag{31}$$

Note that (24) implies that

$$\lim_{w_j \to \infty} |w_j h(w)| \quad \leq \quad \lim_{w_j \to \infty} \left| w_j e^{-\frac{\|w\|^2}{2t}} \right| \quad = \quad 0. \tag{32}$$

Consequently,

$$\int_{\mathbb{R}^d} w_j \frac{\partial}{\partial_{w_j}} h(w)dw \quad = \quad \int_{\mathbb{R}^{d-1}} \left[-\int_{-\infty}^{\infty} h(w)dw_j \right] dw_{-j} \quad = \quad -Z_\delta. \tag{33}$$

Equations (30) and (33) together imply that

$$\mathbb{E}_{\rho_\delta} \left(W_j \frac{\partial}{\partial w_j} g(W) \right) \quad = \quad \delta. \tag{34}$$

The linearity of expectations gives (22) completing the proof. ■

Appendix B. HJ-Prox-based PGD Convergence

For completeness and ease of presentation, we restate the theorem.

Proof of Thm. 3. *Let f, g be proper, LSC, and convex, with f additionally L-smooth. Consider the HJ-Prox-based PGD iteration given by*

$$x_{k+1} \quad = \quad \text{prox}_{tg}^{\delta_k} \left(x_k - t\nabla f(x_k) \right), \quad k = 1, \ldots, \tag{35}$$

with step size $0 < t < 2/L$ and $\left\{ \sqrt{\delta_k} \right\}_{k \geq 1}$ a summable sequence. Then x_k converges to a minimizer of $f + g$.

Proof For appropriately chosen step-size t, the PGD algorithm map is averaged and its fixed points coincide with the global minimizers of f (as shown in the Lemma below).

Lemma 7 (Averagedness and Fixed Points of PGD) *Let $0 < t < \frac{2}{L}$ and define, for $x \in \mathbb{R}^n$*

$$T(x) \quad = \quad \operatorname{prox}_{tg}\big(x - t\nabla f(x)\big). \tag{36}$$

Then T is an averaged operator, and its fixed points Fix(T) coincide with f's global minimizers X^* [19, Section 4.2].

The PGD iterates are computed by applying the mapping $T(x) = \operatorname{prox}_{tg}(x - t\nabla f(x))$. By Lemma 7, T is an averaged operator and $x_k \to x^* \in \text{Fix}(T) = X^*$.

The HJ-PGD iterates can be written as

$$\hat{x}_{k+1} \quad = \quad \operatorname{prox}_{tg}^{\delta_k}(\hat{x}_k - t\nabla f(\hat{x}_k)) \quad = \quad T(\hat{x}_k) + \varepsilon_k, \tag{37}$$

where

$$\varepsilon_k \quad = \quad \operatorname{prox}_{tg}^{\delta_k}(\hat{x}_k - t\nabla f(\hat{x}_k)) - \operatorname{prox}_{tg}(\hat{x}_k - t\nabla f(\hat{x}_k)). \tag{38}$$

Since $\sum_k \sqrt{\delta_k}$ is finite, $\sum_k^\infty \|\varepsilon_k\|$ is finite by Theorem 2. Consequently, $\hat{x}_k \to x^\star \in X^*$ by Theorem 1. ■

Appendix C. HJ-Prox-based DRS Convergence

We restate the statement of the theorem for readability.

Proof of Thm. 4. *Let f, g be proper, convex, and LSC. Consider the HJ-Prox–based DRS iteration given by*

$$\begin{aligned} x_{k+1/2} \quad &= \quad \operatorname{prox}_{tf}^{\delta_k}(z_k), \\ x_{k+1} \quad &= \quad \operatorname{prox}_{tg}^{\delta_k}(2x_{k+1/2} - z_k), \\ z_{k+1} \quad &= \quad z_k + x_{k+1} - x_{k+1/2}, \end{aligned} \tag{39}$$

with $\left\{\sqrt{\delta_k}\right\}_{k\geq 1}$ a summable sequence. Then x_k converges to a minimizer of $f + g$.

Proof The DRS algorithm map is averaged and its fixed points coincide with the global minimizers of f.

Lemma 8 (Averagedness and Fixed Points of DRS) *Let $t > 0$ and define, for $z \in \mathbb{R}^n$*

$$T(z) \quad = \quad z + \operatorname{prox}_{tg}\big(2\ \operatorname{prox}_{tf}(z) - z\big) - \operatorname{prox}_{tf}(z). \tag{40}$$

Note this is the fixed point operator for the dual variable in the DRS algorithm. Then T is firmly nonexpansive (hence averaged), and

$$\mathit{Fix}(T) \quad = \quad \{z : \operatorname{prox}_{tf}(z) \in Z^*\}. \tag{41}$$

*[**?** , Remark 5]*

By Lemma 8, $z_k \to z^*$ and $\text{prox}(z^*) = x^* \in X^*$. We can express the HJ-DRS update in terms of the DRS algorithm map T (40).

$$\hat{z}_{k+1} \quad = \quad T(\hat{z}_k) + \varepsilon_k, \tag{42}$$

where

$$\begin{aligned} \varepsilon_k &= \text{prox}_{th}(w_k + 2\kappa_k) - \text{prox}_{th}(w_k) + \zeta_k - \kappa_k, && (43)\\ w_k &= 2\,\text{prox}_{tg}(\hat{z}_k) - \hat{z}_k, && (44) \end{aligned}$$

and

$$\begin{aligned} \zeta_k &= \text{prox}_{tg}^{\delta_k}(\hat{z}_k) - \text{prox}_{tg}(\hat{z}_k) && (45)\\ \kappa_k &= \text{prox}_{tf}^{\delta_k}(\hat{z}_k) - \text{prox}_{tf}(\hat{z}_k). && (46) \end{aligned}$$

We have the following bound

$$\|\varepsilon_k\| \quad \leq \quad \|w_k + 2\kappa_k - w_k\| + \|\zeta_k\| + \|\kappa_k\| \quad = \quad 3\|\kappa_k\| + \|\zeta_k\|, \tag{47}$$

which follows from the triangle inequality and the fact that proximal mappings are nonexpansive.

Since $\sum_k \sqrt{\delta_k}$ is finite $\sum_k^\infty \|\varepsilon_k\|$ is finite by Theorem 2. Consequently, $\hat{z}_k \to z^\star$ by Theorem 1. Since proximal maps are continuous, $\hat{x}_k = \text{prox}(\hat{z}_k) \to \text{prox}(z^\star) \in X^*$. ■

Appendix D. HJ-Prox-based DYS Convergence

For completeness and ease of presentation, we restate the theorem.

Proof of Thm. 5. *For DYS, consider $f + g + h$. Let f, g, h be proper, LSC, and convex, with h additionally L-smooth. Consider the HJ-Prox–based DYS algorithm given by*

$$\begin{aligned} y_{k+1} &= \text{prox}_{tf}^{\delta_k}(x_k), \\ z_{k+1} &= \text{prox}_{tg}^{\delta_k}\big(2y_{k+1} - x_k - t\nabla h(y_{k+1})\big) \\ x_{k+1} &= x_k + z_{k+1} - y_{k+1} \end{aligned} \tag{48}$$

with $\{\sqrt{\delta_k}\}_{k\geq 1}$ a summable sequence, and $0 < t < 2/L$. Then x_k converges to a minimizer of $f + g + h$.

Proof For appropriately chosen step-size t, the DYS algorithm map is averaged and its fixed points coincide with the global minimizes of $f + g + h$.

Lemma 9 (Averagedness and Fixed Points of DYS) *Let $t > 0$ and define, for $z \in \mathbb{R}^n$,*

$$T(z) = z - \text{prox}_{tf}(z) + \text{prox}_{tg}\big(2\,\text{prox}_{tf}(z) - z - t\nabla h(\text{prox}_{tf}(z))\big). \tag{49}$$

Note this is the fixed point operator for the DYS algorithm and its fixed points Fix(T) coincide with global minimizers X^*. T is firmly nonexpansive (hence averaged), and

$$Fix(T) = \{z : x \in X^\star\}. \tag{50}$$

[7, Theorem 3.1]

By Lemma 9, $z_k \to z^\star$ and $z^* \in X^*$. We can express the HJ-DYS update in terms of DYS algorithm map T (49).

$$\hat{z}_{k+1} \quad = \quad T(\hat{z}_k) + \varepsilon_k, \tag{51}$$

where

$$\begin{aligned} \varepsilon_k &= \operatorname{prox}_{tg}\left(S_t(z_k) + d_k\right) - \operatorname{prox}_{tg}\left(S_t(z_k)\right) + \zeta_k - \kappa_k && (52)\\ S_t(z_k) &= 2\operatorname{prox}_{tf}(z_k) - z_k - t\nabla h\big(\operatorname{prox}_{tf}(z_k)\big) && (53)\\ d_k &= 2\kappa_k - t[\nabla h\big(\operatorname{prox}_{tf}(z_k) + \kappa_k\big) - \nabla h\big(\operatorname{prox}_{tg}(z_k)\big)] && (54) \end{aligned}$$

and

$$\begin{aligned} \zeta_k &= \operatorname{prox}_{tg}^{\delta_k}(\hat{z}_k) - \operatorname{prox}_{tg}(\hat{z}_k) && (55)\\ \kappa_k &= \operatorname{prox}_{tf}^{\delta_k}(\hat{z}_k) - \operatorname{prox}_{tf}(\hat{z}_k). && (56) \end{aligned}$$

We have the following bound

$$\|\varepsilon_k\| \quad \leq \quad (1+tL)\|\kappa_k\| + \|\zeta_k\|, \tag{57}$$

which follows from the triangle inequality, L-smoothness of h, and the fact that proximal mappings are nonexpansive.

Since $\sum_k \sqrt{\delta_k}$ is finite $\sum_k^\infty \|\varepsilon_k\|$ is finite by Theorem 2. Consequently, $\hat{z}_k \to z^\star$ where $z^\star$ is a global minimizer of $f+g+h$ by Theorem 1 and Lemma 9. ■

Appendix E. HJ-Prox-based PDHG Convergence

For completeness and ease of presentation, we restate the theorem.

Proof of Thm. 6. *Let f, g be proper, convex, and LSC. Consider the HJ-Prox–based PDHG algorithm given by*

$$\begin{aligned} y_{k+1} &= \operatorname{prox}_{\sigma g^*}^{\delta_k}(y_k + \sigma A x_k),\\ x_{k+1} &= \operatorname{prox}_{\tau f}^{\delta_k}(x_k - \tau A^\top y_{k+1}), \end{aligned} \tag{58}$$

with parameters $\tau, \sigma > 0$ satisfying $\tau\sigma\|A\|^2 < 1$ and $\{\sqrt{\delta_k}\}_{k\geq 1}$ a summable sequence. Where g^* denotes the Fenchel conjugate of g. Then x^k converges to a minimizer of $f(x) + g(Ax)$.

Proof For appropriately chosen τ, σ the PDHG algorithm map is averaged and its fixed points corresponding to x_k updates coincide with the global minimizes of $f(x) + g(Ax)$.

Lemma 10 (Averagedness and Fixed Points of PDHG) *Let $\tau, \sigma > 0$ satisfying $\tau\sigma\|A\|^2 < 1$ and define, for $z \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$*

$$T(z,w) = \begin{bmatrix} \operatorname{prox}_{\tau f}\big(z - \tau A^\top \operatorname{prox}_{\sigma g^*}(w + \sigma A z)\big) \\ \operatorname{prox}_{\sigma g^*}(w + \sigma A z) \end{bmatrix}. \tag{59}$$

Let $V = \operatorname{diag}(\frac{1}{\tau} I_n, \frac{1}{\sigma} I_m)$. On a product space with a weighted inner product $\langle (x,y),(x',y')\rangle_V = \frac{1}{\tau}\langle x, x'\rangle + \frac{1}{\sigma}\langle y, y'\rangle$, the map T is an averaged operator. Note this is the fixed point operator for the

PDHS algorithm and its fixed points Fix(T) coincide with the set of primal-dual KKT saddle points for $f(x) + g(Ax)$, where the primal point coincides with the global minimizers X^*. T is firmly nonexpansive (hence averaged), and

$$Fix(T) = \{(z^*, w^*) : z^* \in X^\star\}. \tag{60}$$

[3, Algorithm 1, Thm. 1] [9, Lemma 2]

By Lemma 10, $z_k \to z^\star$ and $z^* \in X^*$. We can express the HJ-PDHG update in terms of PDHG algorithm map T (59).

$$(\hat{z}_{k+1}, \hat{w}_{k+1}) \quad = \quad T(\hat{z}_k, \hat{w}_k) + \varepsilon_k \tag{61}$$

where

$$\varepsilon_k \quad = \quad \begin{bmatrix} \mathrm{prox}_{\tau f}(u_k - \tau A^\top \zeta_k) - \mathrm{prox}_{\tau f}(u_k) + \kappa_k \\ \zeta_k \end{bmatrix} \tag{62}$$

$$u_k \quad = \quad \hat{z}_k - \tau A^\top \mathrm{prox}_{\sigma g^\star}(w_k + \sigma A \hat{z}_k), \tag{63}$$

and

$$\zeta_k \quad = \quad \mathrm{prox}^{\delta_k}_{\sigma g^*}(\hat{z}_k + \sigma A \hat{w}_k) - \mathrm{prox}_{\sigma g^*}(\hat{z}_k + \sigma A \hat{w}_k) \tag{64}$$

$$\kappa_k \quad = \quad \mathrm{prox}^{\delta_k}_{\tau f}(\hat{w}_k - \tau A^\top \hat{z}_k) - \mathrm{prox}_{\tau f}(\hat{w}_k - \tau A^\top \hat{z}_k) \tag{65}$$

In the weighted norm $\|(w, z)\|_V^2 = \frac{1}{\tau}\|w\|^2 + \frac{1}{\sigma}\|z\|^2$, we have the following bound

$$\|\varepsilon_k\|_V^2 \quad \leq \quad \big(2\tau \|A\|_{\mathrm{op}}^2 + \frac{1}{\sigma}\big)\|\zeta_k\|^2 + \frac{2}{\tau}\|\kappa_k\|^2 \tag{66}$$

which follows from the fact that proximal mappings are nonexpansive and from $\|A^\top\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}$. Since $\sum_k \sqrt{\delta_k}$ is finite $\sum_k^\infty \|\varepsilon_k\|$ is finite by Theorem 2. Consequently, $(z_k, w_k) \to (z^*, w^*)$ where $z^\star$ is a global minimizer of $f + g$ by Theorem 1 and Lemma 10. ■

Appendix F. Experiment Details

HJ-Prox and analytical counterparts run through all iterations. Every experiment simulates a ground truth structure with added noise and blur depending on problem setup. All parameters and step sizes are matched between HJ-Prox and the analytical counterparts to ensure a fair comparison. The HJ-Prox δ sequence follows a schedule

$$\delta_k \quad = \quad \frac{1}{k^{2.00001}}, \tag{67}$$

where k denotes iteration number. The defined schedule decays strictly faster than $1/k^2$ satisfying conditions used in Theorem 1.

F.1. PGD: LASSO Regression

We solve the classic LASSO regression problem using PGD. The simulation setup involves a design matrix $X \in \mathbb{R}^{250\times500}$ with 250 observations and 500 predictors. The true coefficients β are set such that $\beta^{400:410} = 1$ and all others are zero. The objective function is written as,

$$\arg\min_{\beta} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1 \tag{68}$$

$$X \in \mathbb{R}^{250\times500}, \quad \beta \in \mathbb{R}^{500}, \quad y \in \mathbb{R}^{250}.$$

The analytical PGD baseline performs a gradient step on the least-squares term followed by the exact soft thresholding.

F.2. DRS: Multitask Regression

Multitask regression learns predictive models for multiple related response variables by sharing information across tasks to enhance performance. We solve this problem using Douglas-Rachford splitting, employing HJ-Prox in place of analytical updates. We group the quadratic loss with the nuclear norm regularizer to form one function and the row and column group LASSO terms to form the other. Both resulting functions are non-smooth, requiring HJ-Prox for their proximal mappings. The simulation setup involves $n = 50$ observations, $p = 30$ predictors, and $q = 9$ tasks. The objective function is written as,

$$\arg\min_{B} \frac{1}{2}\|XB - Y\|_F^2 + \lambda_1\|B\|_* + \lambda_2\sum_i \|b_{i,\cdot}\|_2 + \lambda_3\sum_j \|b_{\cdot,j}\|_2 \tag{69}$$

$$X \in \mathbb{R}^{50\times30}, \quad B \in \mathbb{R}^{30\times9}, \quad Y \in \mathbb{R}^{50\times9}.$$

The analytical counterpart for Douglas Rachford Splitting utilizes singular value soft thresholding for the nuclear norm and group soft thresholding for the row and column penalties. These regularizers are integrated with fast iterative soft thresholding (FISTA) to handle the data fidelity term with nuclear norm regularization and Dykstra's algorithm to handle the sum of row and column group LASSO penalties.

F.3. DRS: Fused LASSO

The fused LASSO is commonly used in signal processing to promote piecewise smoothness in the solution. We apply it to recover a Doppler signal with length $n = 256$ using a third-order differencing matrix D. We solve this problem with DRS, comparing two implementation strategies: an exact method using product-space reformulation motivated by [23], and an approximate method using HJ-Prox. The objective function is written as,

$$\arg\min_{B} \frac{1}{2}\|\beta - y\|^2 + \lambda\|D\beta\|_1 \tag{70}$$

$$\beta \in \mathbb{R}^{256}, \quad y \in \mathbb{R}^{256}, \quad D \in \mathbb{R}^{253\times256}.$$

The proximal operator of $\lambda\|D\beta\|_1$ has no closed-form solution for general linear operators D. The analytical counterpart addresses this by reformulating the problem in a product space with

auxiliary variable $w = D\beta$, yielding separable proximal operators (weighted averaging and soft thresholding) at the cost of inverting terms including $D^\top D$ at each iteration. In contrast, our HJ-Prox variant directly approximates the intractable proximal operator through Monte Carlo sampling. As a reminder, both use identical DRS parameters for fair comparison.

F.4. DYS: Sparse Group LASSO

The sparse group LASSO promotes group-level sparsity while allowing individual variable selection within groups, which is useful when certain groups are relevant but contain unnecessary variables. We solve this problem using DYS, employing HJ-Prox for the proximal operators of the non-smooth regularizers. The simulation setup involves $n = 300$ observations with $G = 6$ groups, each having 10 predictors. The objective function is written as,

$$\arg\min_{\beta} \frac{1}{2} \|X\beta - y\|_2^2 \; + \; \lambda_1 \sum_{g=1}^{G} \|\beta_g\|_2 \; + \; \lambda_2 \|\beta\|_1 \tag{71}$$

$$X \in \mathbb{R}^{300\times 60}, \quad \beta \in \mathbb{R}^{60}, \quad y \in \mathbb{R}^{300}.$$

The analytical counterpart for DYS solves the sparse group LASSO by using soft thresholding for the ℓ_1 penalty and group soft thresholding for the group ℓ_2 penalty.

F.5. PDHG: Total Variation

Lastly, we implement PDHG method to solve the isotropic total variation regularized least-squares problem. We apply the proximal operator for the data fidelity term via its closed-form update and employ our HJ-based proximal operator for the total variation penalty. For this experiment, we recover a smoothed 64 x 64 black and white image from a noisy and blurred image y. The objective function is written as,

$$\arg\min_{\beta} \frac{1}{2} \|X\beta - y\|_F^2 + \lambda \mathrm{TV}(\beta) \tag{72}$$

$$\beta \in \mathbb{R}^{64\times 64}, \quad y \in \mathbb{R}^{64\times 64}.$$

The (slightly smoothed) isotropic TV we use to evaluate the objective is

$$\mathrm{TV}(\beta) \;\; = \;\; \sum_{i=1}^{64} \sum_{j=1}^{64} \sqrt{(\nabla_x \beta)_{i,j}^2 + (\nabla_y \beta)_{i,j}^2}. \tag{73}$$

The analytical counterpart for PDHG algorithm updates dual variables using closed-form scaling for data fidelity and clamping (for ℓ_2 projection of TV dual), and primal variables using Fast Fourier transform convolution and divergence via finite differences.