Stochastic model-based minimization of weakly convex functions

Fengjing Zhu
Institutes of Science and Development
Chinese Academy of Sciences
Beijing
zhufengjing24@mails.ucas.ac.cn

Abstract

This report reviews the paper *Stochastic Model-Based Minimization of Weakly Convex Functions*, which develops a general framework for minimizing weakly convex and possibly nonsmooth functions via stochastic models. The proposed algorithm iteratively samples a stochastic one-sided model of the objective and performs a proximal minimization step. Under standard Lipschitz and weak convexity assumptions, the authors prove that the expected norm of the gradient of the Moreau envelope converges at a rate of $\mathcal{O}(k^{-1/4})$. The framework unifies and establishes complexity guarantees for several key algorithms, including the stochastic proximal point, proximal subgradient, and prox-linear methods. This report summarizes the algorithmic structure, theoretical analysis, and practical implications of the results. For code and reproducibility, see: `link`.

1 Introduction

Optimization is fundamental in machine learning, signal processing, and operations research. Many real-world problems, such as robust regression and neural network training, are nonconvex and nonsmooth, making classical optimization techniques inadequate.

A key class of interest is *weakly convex functions*, which allow convexity after adding a quadratic term. To address stochastic minimization under such settings, Davis and Drusvyatskiy propose a unified stochastic model-based framework. This approach replaces exact gradients with one-sided model approximations and introduces regularization via proximal steps.

Their algorithm covers stochastic proximal point, prox-linear, and subgradient methods, and evaluates stationarity via the Moreau envelope. It achieves the first $\mathcal{O}(k^{-1/4})$ convergence guarantee for this broad problem class.

The overall framework is summarized in Figure 1.

2 Related Works

Nonsmooth, nonconvex optimization problems arise in diverse domains such as phase retrieval, covariance estimation, and neural network training. Classical methods often struggle due to non-differentiability and non-convexity. Recent advancesincluding subgradient methods (Davis et al. 2018), proximal algorithms (Duchi and Ruan 2019), and convex relaxations (Chen et al. 2015)offer partial remedies, but typically depend on structural assumptions or careful initialization.

The *Moreau envelope* plays a central role in analyzing such problems, providing a smooth surrogate whose gradient quantifies stationarity. Davis and Drusvyatskiy leveraged this in a stochastic model-based framework, establishing the first $\mathcal{O}(k^{-1/4})$ convergence rate for weakly convex functions.

Problem Definition

Minimize $\varphi(x) = g(x) + r(x)$ with stochastic model $g_x(\cdot, \xi)$ and closed $r(x)$

Problem Property

Function $\varphi(x)$ is $(\tau + \eta)$-weakly convex

Algorithm: Model-Based Minimization

Sample $\xi_t \sim \mathbb{P}$

Minimize $r(x) + g_{x_t}(x, \xi_t) + \frac{\beta_t}{2}\|x - x_t\|^2$

Return x_{t^*} with probability $\propto \frac{\hat{\rho}}{\beta_t - \tau - \eta}$

Stationarity Measure

Use $\nabla \varphi_{1/\hat{\rho}}(x)$ with $\mathcal{O}(k^{-1/4})$ guarantee

Applications

Stochastic Proximal Point, Prox-Linear, Subgradient

Figure 1: Overview: from problem setup to algorithm, analysis, and applications.

Extensions of the envelope to *Bregman distances* (Kan and Song 2012; Bauschke et al. 2022) have broadened its use in variational analysis and imaging (Durmus et al. 2018), though most results remain limited to convex settings.

In application, *Nonnegative Matrix Factorization (NMF)* (Gillis 2017) exemplifies constrained non-convex problems, but remains challenging under noise and incomplete observations.

3 Main Results

3.1 Problem Setting

We consider the stochastic composite optimization problem:

$$\min_{x \in \mathbb{R}^d} \ \varphi(x) := g(x) + r(x),$$

where $g : \mathbb{R}^d \to \mathbb{R}$ is possibly nonconvex and nonsmooth, and $r : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a proper closed function, often used for regularization or constraints.

The algorithm accesses g via a stochastic one-sided model $g_x(y, \xi)$, where $\xi \sim \mathbb{P}$. We assume:

(A1) i.i.d. samples $\xi_t \sim \mathbb{P}$ are available;

(A2) $\mathbb{E}_\xi[g_x(x, \xi)] = g(x)$, and $\mathbb{E}_\xi[g_x(y, \xi)] \geq g(y) - \frac{\tau}{2}\|y - x\|^2$;

(A3) $y \mapsto g_x(y, \xi) + r(y)$ is η-weakly convex;

(A4) g and $g_x(\cdot, \xi)$ are L-Lipschitz in expectation.

Then φ is $(\tau + \eta)$-weakly convex, satisfying:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y) + \frac{\rho \lambda (1 - \lambda)}{2}\|x - y\|^2.$$

3.2 Algorithm Description

The method follows a stochastic model-based minimization framework, where at each step, a model of the objective is sampled and minimized with a proximal term. The procedure is shown in Algorithm 1.

Algorithm 1: Stochastic Model Based Minimization

Input: $x_0 \in \mathbb{R}^d$, parameters $\hat{\rho} > \tau + \eta$, sequence $\{\beta_t\}_{t=0}^{T} \subset (\hat{\rho}, \infty)$, iteration count T

for $t = 0$ **to** T **do**

  Sample $\xi_t \sim \mathbb{P}$;

  Set $x_{t+1} = \arg\min_{x} \left\{ r(x) + g_{x_t}(x, \xi_t) + \frac{\beta_t}{2}\|x - x_t\|^2 \right\}$;

Sample $t^* \in \{0, \dots, T\}$ with probability $\mathbb{P}(t^* = t) \propto \frac{\hat{\rho} - \tau - \eta}{\beta_t - \eta}$;

Output: x_{t^*}

3.3 Stationarity Measure via Moreau Envelope

To evaluate convergence, we use the *Moreau envelope*:

$$\varphi_\lambda(x) := \inf_y \left\{ \varphi(y) + \frac{1}{2\lambda}\|y - x\|^2 \right\},$$

which is differentiable if φ is ρ-weakly convex and $\lambda < 1/\rho$. Its gradient,

$$\nabla \varphi_\lambda(x) = \frac{1}{\lambda}(x - \text{prox}_{\lambda\varphi}(x)),$$

serves as a continuous stationarity measure.

Main Result. Under (A1)(A4), the output x_{t^*} satisfies:

$$\mathbb{E}\left[\|\nabla \varphi_{1/\hat{\rho}}(x_{t^*})\|^2\right] \leq \frac{\hat{\rho}(\varphi_{1/\hat{\rho}}(x_0) - \min \varphi) + 2\hat{\rho}^2 L^2 \sum_{t=0}^{T} \frac{1}{(\beta_t - \eta)(\beta_t - \hat{\rho})}}{\sum_{t=0}^{T} \frac{2(\hat{\rho} - \tau - \eta)}{\beta_t - \eta}}.$$

Convergence Rate. Choosing $\beta_t = \hat{\rho} + \frac{1}{\alpha\sqrt{T+1}}$ yields

$$\mathbb{E}\left[\|\nabla \varphi_{1/\hat{\rho}}(x_{t^*})\|^2\right] = \mathcal{O}(T^{-1/2}),$$

implying complexity $\mathcal{O}(\varepsilon^{-2})$ for ε-stationarity.

Proof Sketch. The proof applies Lemma 3.1 to bound the envelope value at x_{t+1}:

$$\mathbb{E}_t[\varphi_{1/\hat{\rho}}(x_{t+1})] \leq \varphi_{1/\hat{\rho}}(x_t) - \frac{\hat{\rho}(\hat{\rho} - \tau - \eta)}{2(\beta_t - \eta)}\|x_t - \hat{x}_t\|^2 + \frac{2\hat{\rho}^2 L^2}{(\beta_t - \eta)(\beta_t - \hat{\rho})}.$$

Summing over t and using $\varphi_{1/\hat{\rho}}(x_{t+1}) \geq \min \varphi$ leads to:

$$\mathbb{E}\left[\sum_{t=0}^{T} \frac{\hat{\rho} - \tau - \eta}{\beta_t - \eta}\|x_t - \hat{x}_t\|^2\right] \leq 2 \cdot \frac{\varphi_{1/\hat{\rho}}(x_0) - \min \varphi}{\hat{\rho}} + 4L^2 \sum_{t=0}^{T} \frac{1}{(\beta_t - \eta)(\beta_t - \hat{\rho})}.$$

Since $\nabla \varphi_{1/\hat{\rho}}(x_t) = \hat{\rho}(x_t - \hat{x}_t)$, we obtain the final bound.

3.4 Special Cases

The framework recovers several classical methods by varying model type and assumptions, summarized in Table 1.

Table 1: Special Cases under the General Framework

Method	Model Type	Key Assumptions
Stochastic Proximal Point	Exact	$\tau = 0$, $\eta = \rho$
Proximal Subgradient	Linear	Bounded variance
Prox-Linear	Composite	Smooth c, convex h

3.5 Numerical Experiment

We consider the stochastic logistic regression problem:

$$\min_{x \in \mathbb{R}^d} \varphi(x) := g(x) + r(x),$$

where $g(x) = \mathbb{E}_\xi \left[\log\left(1 + \exp(-yz^\top x)\right)\right]$ with $\xi = (z, y)$, $y \in \{-1, 1\}$, and $r(x) = \frac{\lambda}{2}\|x\|^2$ with $\lambda = 0.1$.

We initialize at $x_0 = 0$ and run for $T = 100$ iterations, setting $\tau + \eta = 2, \quad \hat{\rho} = 3, \quad \beta_t > \hat{\rho}$.

The experiment is repeated 5 times to evaluate robustness. Results are shown in Figure 2 and 3.

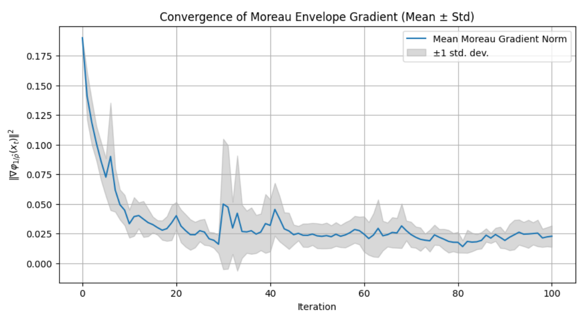

Figure 2: Convergence of Moreau Envelope Gradient

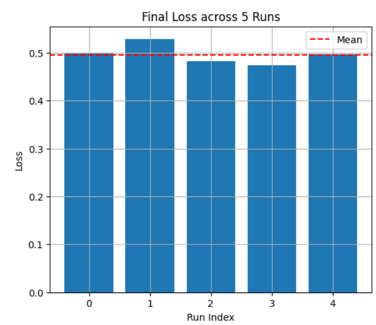

Figure 3: Final Loss across 5 Runs

4 Conclusion

This report examined the stochastic model-based minimization framework for weakly convex functions proposed by Davis and Drusvyatskiy. Its main contribution is a unified algorithmic scheme that generalizes several classical methods and provides the first complexity guarantees for convergence to approximate stationary points. A central analytical tool is the Moreau envelope, whose gradient yields a smooth measure of near-stationarity.

The framework stands out for its generality and minimal assumptions: it requires only stochastic one-sided models accurate in expectation, making it widely applicable. It achieves an $\mathcal{O}(k^{-1/4})$ convergence rate in the Moreau envelope gradient norm, improving known results in certain settings.

However, limitations remain. The theory focuses on first-order stationarity, with second-order or global analysis still unexplored. Performance may be sensitive to hyperparameters such as β_t and $\hat{\rho}$, suggesting the value of adaptive tuning. Future enhancements could address heavy-tailed noise, variance reduction, or structured modeling (e.g., Hessian surrogates) to improve robustness.

Promising directions include applications to large-scale nonconvex learning problemssuch as deep networks and stochastic reinforcement learningwhere this framework's combination of flexibility and theoretical guarantees offers a compelling foundation for stochastic nonsmooth optimization.

5 Some possible readings

Attouch, H (1984). “Variational convergence for functions and operators”. *Applicable Mathematics Series*.

Bauschke, H. H., M. D., and S. Lindstrom (2022). “Using BregmanMoreau envelopes for regularization”. *Set-Valued and Variational Analysis*.

Chen, Y., Y. Chi, and A. J. Goldsmith (2015). “Exact and stable covariance estimation from quadratic sampling via convex programming”. *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059.

Davis, D., D. Drusvyatskiy, K. J. MacPhee, and C. Paquette (2018). “Subgradient methods for sharp weakly convex functions”. *Journal of Optimization Theory and Applications*, vol. 179, no. 3, pp. 962–982.

Davis, D., D. Drusvyatskiy, and C. Paquette (2020). “The nonsmooth landscape of phase retrieval”. *IMA Journal of Numerical Analysis*, vol. 40, no. 3, pp. 2652–2695.

Duchi, J. C. and F. Ruan (2019). “Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval”. *Information and Inference: A Journal of the IMA*, vol. 8, no. 3, pp. 471–529.

Durmus, A., É. Moulines, and M. Pereyra (2018). “Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau”. *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 473–506.

Gillis, N. (2017). “Introduction to Nonnegative Matrix Factorization”. *SIAG/OPT Views and News*, vol. 25, no. 1. arXiv:1703.00663, pp. 7–16.

Kan, I and Y Song (2012). “BregmanMoreau envelopes and Bregman proximity operators”. *Journal of Optimization Theory and Applications*, vol. 154, no. 3, pp. 904–930.

Kecis, I. and L. Thibault (2015). “Moreau envelopes of s-lower regular functions”. *Nonlinear Analysis: Theory, Methods & Applications*, vol. 127, pp. 157–181.

Moreau, J.-J. (1965). “Proximité et dualité dans un espace hilbertien”. *Bulletin de la Société Mathématique de France*, vol. 93, pp. 273–299.

Pérez-Aros, P. and E. Vilches (2021). “Moreau envelope of supremum functions with applications to infinite and stochastic programming”. *SIAM Journal on Optimization*, vol. 31, no. 3, pp. 1635–1657.

Acknowledgement

I would like to express my sincere gratitude to Professor Yaoliang Yu for his inspiring lectures and invaluable guidance throughout the course. His clear and accessible teaching style, along with the concise and well-structured lecture notes, made complex concepts in optimization theory and stochastic algorithms much more approachable. Thanks to his instruction, I have taken my first solid step into the world of optimization algorithms in data science.