SPEEDING UP NONSMOOTH BAYESIAN MCMC SAMPLING VIA INEXACT PROXIMAL UNADJUSTED LANGEVIN ALGORITHM

Anonymous authors
Paper under double-blind review

ABSTRACT

We study sampling from posterior distributions with nonsmooth composite potentials, a setting in which proximal-based Langevin methods are theoretically appealing but in practice limited to simple functions with closed-form proximal operators. We introduce iPULA for composite potentials, an inexact proximal unadjusted Langevin algorithm that replaces exact proximal steps with controlled approximations. Our approach leverages the Moreau envelope to smooth the potential, while allowing inexact evaluation of its gradient through inexact proximal computations. Our main theoretical result establishes a non-asymptotic analysis of iPULA with biased and potentially history-dependent inexact drifts. Under convexity of the potential, we derive an explicit Kullback–Leibler bound for the averaged iPULA law, showing that square-summable oracle errors do not introduce an additional asymptotic error term beyond the standard constant-step ULA discretization floor. Under the Polyak–Łojasiewicz condition, we further obtain bounds on the objective gap, while under strong convexity, we establish geometric Wasserstein guarantees for the final iterate. We demonstrate the practical relevance of iPULA on image reconstruction tasks, where proximal operators cannot be computed exactly. Experiments demonstrate the effectiveness of iPULA and support our theoretical results.

1 INTRODUCTION

Markov chain Monte Carlo (MCMC) methods provide a foundational framework for posterior inference in Bayesian modeling when normalizing constants are analytically intractable (Robert & Casella, 2004). In high-dimensional regimes, standard random-walk algorithms suffer from severe autocorrelation and sluggish mixing, requiring excessively small step-sizes to maintain viable acceptance rates (Brooks et al., 2011). This limitation motivates gradient-based MCMC algorithms such as the unadjusted Langevin algorithm (ULA), which exploit first-order geometry to steer proposals toward high-probability regions and accelerate state-space exploration (Dalalyan, 2017; Dalalyan & Karagulyan, 2019; Durmus et al., 2018; Ma et al., 2015).

While MCMC methods are asymptotically valid, their efficiency often degrades in high dimensions, where random-walk Metropolis exhibits slow mixing and strong sample correlation. As dimensionality increases, maintaining reasonable acceptance rates requires shrinking the proposal step-size, leading to inefficient exploration due to poor adaptation to the target distribution's local structure and high computational costs (Brooks et al., 2011). This motivates gradient-based MCMC methods, such as Langevin dynamics, which use first-order derivative information to guide proposals toward more probable regions (Dalalyan, 2017; Dalalyan & Karagulyan, 2019; Dalalyan et al., 2022; Durmus & Moulines, 2019; Durmus et al., 2018; Ma et al., 2015).By leveraging gradients, these methods adapt to the curvature of the target distribution, enabling larger, more informed moves that significantly improve convergence rates and sampling efficiency. Consequently, Langevin-based approaches offer a principled balance between exploration and exploitation, making them especially important for scalable inference in modern high-dimensional applications (Dubey et al., 2016)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 Related work: Nonsmooth MCMC methods and key question

Nonsmooth potentials arise naturally in applications involving sparsity-inducing priors (e.g., ℓ_1 penalties in compressed sensing and high-dimensional regression), image reconstruction with total variation regularization, nuclear-norm based low-rank matrix estimation, and Bayesian formulations of machine learning models with hinge or absolute-value losses (e.g., Bayesian trendfiltering) (Crucinio et al., 2025; Durmus et al., 2018; Lau & Liu, 2022; Lau et al., 2024; Pereyra, 2016; Salim et al., 2019; Shukla et al., 2025). While gradient-based MCMC methods are not directly applicable in the nonsmooth setting, the need for efficient sampling in high-dimensional regimes remains a need. In this setting, unadjusted Langevin algorithm (ULA) and its variants can be extended either by replacing gradients with subgradients (Habring et al., 2024) or by introducing smooth surrogates of the potential. When coupled with the unadjusted Langevin algorithm (ULA), such smoothing strategies, including Gaussian smoothing (Chatterji et al., 2020; Nesterov & Spokoiny, 2017), the Moreau envelope (Durmus et al., 2018; Kabgani & Ahookhosh, 2025b; 2026a; 2025a; Klatzer et al., 2024; Moreau, 1963; Pereyra, 2016; Planiden & Wang, 2016), and forward–backward envelopes (Ahookhosh et al., 2021; Eftekhari et al., 2023; Ghaderi et al., 2024; Kabgani & Ahookhosh, 2026b; Stella et al., 2017; Themelis et al., 2019), yield well-defined stochastic dynamics that preserve key structural properties of the original objective while enabling efficient gradient-based exploration.

Owing to their computational efficiency, smoothing-based ULA methods provide a powerful and scalable framework for Bayesian inference in nonsmooth models, striking a balance between theoretical tractability and practical performance across a wide range of modern applications. The principal smoothing techniques integrated with ULA are as follows:

- **Moreau envelope** is a smoothing technique that associates to a given function $\mathcal{U}$ a differentiable approximation obtained via infimal convolution with a scaled quadratic function, i.e.,

$$\mathcal{U}_\gamma(x) := \inf_{y\in\mathbb{R}^n}\left\{\mathcal{U}(y) + \frac{1}{2\gamma}\|x-y\|^2\right\}, \tag{1}$$

  thereby preserving key properties such as convexity while it is smooth with the gradient

$$\nabla\mathcal{U}_\gamma(x) = \frac{1}{\gamma}\big(x - \mathbf{Prox}_{\gamma\mathcal{U}}(x)\big), \tag{2}$$

  where $\gamma > 0$ and the *proximal operator* is given by

$$\mathbf{Prox}_{\gamma\mathcal{U}}(x) := \mathop{\mathbf{arg\,min}}_{y\in\mathbb{R}^n}\left\{\mathcal{U}(y) + \frac{1}{2\gamma}\|x-y\|^2\right\}. \tag{3}$$

- **Forward–backward envelope** (FBE) is a smoothing technique that offers an efficient way to generate a differentiable approximation of composite potentials of the form $\mathcal{U}(x) = f(x) + g(x)$, where f is twice continuously differentiable with L-Lipschitz continuous gradient with some $L > 0$ and g is simple with a closed-form proximal operator, i.e.,

$$\mathcal{U}_\gamma^{f,g}(x) := \inf_{y\in\mathbb{R}^n}\left\{f(x) + \langle\nabla f(x), y-x\rangle + g(y) + \frac{1}{2\gamma}\|x-y\|^2\right\}, \tag{4}$$

  which is smooth for $\gamma \in (0, \frac{1}{L}]$ and its gradient is given by

$$\nabla\mathcal{U}_\gamma^{f,g}(x) = \frac{1}{\gamma}\left(I - \gamma\nabla^2 f(x)\right)\left(x - \mathbf{Prox}_{\gamma g}(x - \gamma\nabla f(x))\right). \tag{5}$$

Inexact proximal evaluations have recently been explored in proximal sampling. In particular, Ehrhardt et al. (2024) analyze a proximal Langevin method where it keeps the smooth function f and replaces g by its Moreau envelope. This may alter the set of critical points of the potential function, causing the region of maximum probability in the corresponding posterior distribution to differ from that of the original posterior distribution. By contrast, iPULA computes an approximate proximal step over the *full composite objective* $\mathcal{U} = f + g$, constructing a joint Moreau-gradient oracle that regularizes non-Gaussian posterior geometry. Furthermore, our analysis extends the Wasserstein free-energy framework of Durmus et al. (2019) to history-dependent inexactness, moving beyond standard perturbation bounds (Dalalyan & Karagulyan, 2019; Huggins & Zou, 2017).

Limitations of existing smoothing techniques. While the Moreau envelope and forward–backward envelope (FBE) preserve the minimizers of the original potential, they exhibit several limitations. For the *Moreau envelope*, the proximal operator admits a closed form only when the potential has a particularly simple structure, which is typically not satisfied in composite settings; as a result, its use within ULA is largely restricted to simple models. The applicability of *forward-backward envelope* is further constrained by more stringent requirements: (i) evaluating its gradient involves the Hessian of f, which is computationally prohibitive in high dimensions; (ii) the assumption that f is L-smooth restricts the class of admissible models, excluding many problems with nonsmooth or relatively smooth data-fidelity terms; and (iii) it relies on the simplicity of g to ensure a tractable proximal operator, thereby excluding complex or nonseparable regularizers. Collectively, these limitations hinder the practicality of FBE-based ULA methods for many applications. Accordingly, a natural key question arises in the community:

"*How can we design Moreau-smoothed unadjusted Langevin methods for nonsmooth composite potentials when the proximal mapping of the full potential is not available in closed form?*"

1.2 Contribution

To address this question, we first approximate the proximal operator in (3), which in turn yields an inexact oracle for the Moreau envelope in (1). We then incorporate this inexact gradient into the unadjusted Langevin algorithm (ULA), resulting in an inexact proximal-based ULA scheme. The key elements of this construction are summarized below and form the main contributions of this paper:

- (**Inexact oracle of Moreau envelope for ULA.**) In the composite setting, we first solve the proximal subproblem (3) approximately and substitute the result into (2), thereby obtaining an inexact gradient of the Moreau envelope (1); see Section 2.2. Notably, this inexact approach removes the need for $\mathcal{U}$ to admit a simple structure, in contrast to the exact Moreau and forward–backward envelope constructions discussed in Section 1.1. We then incorporate this inexact gradient into the ULA update, yielding the inexact proximal unadjusted Langevin algorithm (iPULA). The resulting framework enables approximate computation of the proximal mapping of $\mathcal{U}$, in contrast to Ehrhardt et al. (2024), which considers an inexact proximal mapping of g, and provides a computable certificate for controlling the resulting Moreau-gradient error. To the best of our knowledge, this is the first proximal-based ULA method that leverages an inexact oracle for the full objective $\mathcal{U}$ to construct the Moreau envelope of composite potentials arising in Bayesian learning.
- (**Non-asymptotic convergence analysis of iPULA.**) Assuming convexity of the potential, we establish an explicit non-asymptotic Kullback–Leibler bound for the averaged iPULA law in the presence of biased, potentially history-dependent oracle errors. In particular, square-summable oracle errors introduce no additional asymptotic term beyond the usual constant-step ULA discretization floor. Under an additional Polyak–Łojasiewicz inequality, we obtain objective-gap bounds, while strong convexity yields geometric final-iterate Wasserstein guarantees; see Section 3.

1.3 Notation

Throughout the paper, $\mathbb{R}^n$ denotes the n-dimensional Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$; $\overline{\mathbb{R}} := \mathbb{R}\cup\{+\infty\}$; $\mathbb{N}_0 := \mathbb{N}\cup\{0\}$. For a proper function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$, $\mathbf{dom}(h) := \{x : h(x) < +\infty\}$. $\mathcal{P}_2(\mathbb{R}^n)$ denotes Borel probability measures with finite second moment. For $u, v \in \mathcal{P}_2(\mathbb{R}^n)$, $\mathcal{W}_2(u,v)$ denotes the 2-Wasserstein distance $\mathcal{W}_2^2(u,v) = \mathbf{inf}_{\xi\in\Gamma(u,v)} \int \|x-y\|_2^2 \, d\xi(x,y)$, $D_{\mathrm{KL}}(u\|v) := \int \log(du/dv)\, du$ is the Kullback–Leibler divergence, and $\mathrm{TV}(u,v) = \frac{1}{2}\int |p(x) - q(x)|\, dx$ is the total variation distance for densities p, q.

2 Problem setup and Moreau envelope

Let us consider *sampling* from a canonical target distribution $\pi(x) : \mathbb{R}^n \to \mathbb{R}$ given by

$$\pi(x) := \frac{e^{-\mathcal{U}(x)}}{\int_{\mathbb{R}^n} e^{-\mathcal{U}(y)}\, dy}, \tag{6}$$

where $\mathcal{U} : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper, lower semicontinuous, convex function which is measurable, i.e,

$$0 < \int_{\mathbb{R}^n} e^{-\mathcal{U}(x)}\, dx < +\infty. \tag{7}$$

Bayesian inference is a popular approach for constructing a target distribution, the posterior, over the unknown parameters of a model by combining prior beliefs with the likelihood of the observed data. The prior encodes existing knowledge about the problem and thereby modulates parameter uncertainty. By Bayes' rule,

$$\pi(x \mid D) \propto P(D \mid x)\, P(x), \tag{8}$$

where $D \in \mathbb{R}^{n\times p}$ is the data set, $x \in \mathbb{R}^p$ is the parameter, $P(D \mid x)$ is the likelihood, and $P(x)$ is the prior. The associated potential is then, up to an additive constant, $\mathcal{U}(x) := -\log \pi(x \mid D) = f(x) + g(x)$, with $f(x) := -\log P(D \mid x)$ and $g(x) := -\log P(x)$.

Assumption 2.1 (**Baseline assumptions**). *Throughout the paper, unless otherwise stated, we assume:*

(a) *$\mathcal{U} : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lower semicontinuous, convex;*

(b) *the integrability condition (7) holds.*

2.1 Proximal operator and Moreau envelope

The following standard facts summarize the properties needed in our analysis.

Fact 2.2 (**Properties of the Moreau envelope**). *Bauschke & Combettes (2017); Planiden & Wang (2016) Under Assumption 2.1, for any $\gamma > 0$: (a) $\mathcal{U}_\gamma$ is continuously differentiable with $\nabla \mathcal{U}_\gamma(x) = \frac{1}{\gamma}(x - \mathbf{Prox}_{\gamma\mathcal{U}}(x))$; (b) $\nabla \mathcal{U}_\gamma$ is L_γ-Lipschitz continuous with $L_\gamma = 1/\gamma$.*

Fact 2.3 (**Strong convexity of envelope**). *(Planiden & Wang, 2016, Lemma 2.23) Let $\mathcal{U} : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lower semicontinuous, and convex, and let $\gamma > 0$. Then $\mathcal{U}$ is σ-strongly convex if and only if $\mathcal{U}_\gamma$ is $\frac{\sigma}{1+\gamma\sigma}$-strongly convex.*

2.2 Inexact first-order oracle for Moreau envelope

The gradient formula in Fact 2.2 shows that evaluating $\nabla \mathcal{U}_\gamma(x)$ requires computing the exact proximal point $\mathbf{Prox}_{\gamma\mathcal{U}}(x)$. In many composite Bayesian models, however, the proximal operator of the full potential $\mathcal{U} = f + g$ is not available in closed form. This motivates the use of an approximate proximal point and, consequently, an inexact first-order oracle for the Moreau envelope.

Let $x_k \in \mathbb{R}^n$ be the current point and let $\varepsilon_k > 0$ be a prescribed lower-level tolerance. We denote by $y_k^{\varepsilon_k}$ an approximate solution of the proximal subproblem (1). Using $y_k^{\varepsilon_k}$ in place of the exact proximal point gives the inexact Moreau-gradient oracle

$$\nabla \mathcal{U}_\gamma^{\varepsilon_k}(x_k) := \frac{1}{\gamma}(x_k - y_k^{\varepsilon_k}). \tag{9}$$

We emphasize that $\nabla \mathcal{U}_\gamma^{\varepsilon_k}(x_k)$ is not, in general, the exact gradient of any explicitly defined smooth function. It is an inexact oracle for the exact Moreau-envelope gradient. The induced gradient error is defined by

$$e_k := \nabla \mathcal{U}_\gamma^{\varepsilon_k}(x_k) - \nabla \mathcal{U}_\gamma(x_k), \tag{10}$$

and we set $\delta_k := \|e_k\|$. The quantity δ_k measures the error propagated from the lower-level proximal solve to the upper-level Langevin update. It is useful for the theoretical analysis, but it is not directly computable in practice because it requires knowledge of the exact proximal point $\mathbf{Prox}_{\gamma\mathcal{U}}(x_k)$. *We say $y_k^{\varepsilon_k}$ is acceptable approximation of $\mathbf{Prox}_{\gamma\mathcal{U}}(x_k)$ if $\delta_k \le \varepsilon_k$, for the accuracy parameter $\varepsilon_k > 0$.*

2.3 Sufficient certificate for acceptable inexact proximal term

Here, we introduce three sufficient conditions guaranteeing that $y_k^{\varepsilon_k}$ is an acceptable proximal term.

Condition 1. For an approximate proximal point $y_k^{\varepsilon_k} \in \mathbf{dom}(\partial \mathcal{U})$, we define

$$r_k := \mathbf{dist}\left(0,\, \partial \mathcal{U}(y_k^{\varepsilon_k}) + \frac{1}{\gamma}\,(y_k^{\varepsilon_k} - x_k)\right), \tag{11}$$

motivated by the first-order optimality condition of the proximal subproblem. We also define

$$\widehat{r}_k := \left\| u_k + \frac{1}{\gamma}(y_k^{\varepsilon_k} - x_k) \right\|, \quad u_k \in \partial \mathcal{U}(y_k^{\varepsilon_k}). \tag{12}$$

Proposition 2.4 (Residual-to-gradient error bound). *Let $\mathcal{U} : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lower semi-continuous, and convex, and let $\gamma > 0$. For a given $x_k \in \mathbb{R}^n$, let $y_k = \mathbf{Prox}_{\gamma\mathcal{U}}(x_k)$ and let $y_k^{\varepsilon_k} \in \mathbf{dom}(\partial\mathcal{U})$ be an approximate proximal point with residual r_k defined by (11). Then, for every selected $u_k \in \partial\mathcal{U}(y_k^{\varepsilon_k})$,*

$$\delta_k = \left\| \nabla \mathcal{U}_\gamma^{\varepsilon_k}(x_k) - \nabla \mathcal{U}_\gamma(x_k) \right\| \le r_k \le \widehat{r}_k. \tag{13}$$

In particular, if $\widehat{r}_k \le \varepsilon_k$, then $\delta_k \le \varepsilon_k$.

Condition 2. For a given outer iterate x_k, define the proximal objective

$$F_k(y) := \mathcal{U}(y) + \frac{1}{2\gamma}\|y - x_k\|^2, \tag{14}$$

and let $y_k^\star := \mathbf{arg\,min}_{y \in \mathbb{R}^n} F_k(y)$. Since $\mathcal{U}$ is convex, F_k is $1/\gamma$-strongly convex and $y_k^\star$ is unique. Suppose that, at an inner iteration t, the lower-level solver returns a primal point $y_{k,t}$ together with a computable lower bound $\ell_{k,t}$ satisfying $\ell_{k,t} \le F_k(y_k^\star)$. We define the corresponding computable gap certificate by

$$\mathcal{G}_{k,t} := F_k(y_{k,t}) - \ell_{k,t}. \tag{15}$$

Proposition 2.5 (Gap-to-gradient error bound). *Suppose Assumption 2.1 holds and let $\gamma > 0$. For any $x_k \in \mathbb{R}^n$, let $y_k^\star = \mathbf{Prox}_{\gamma\mathcal{U}}(x_k)$. If an approximate proximal point $y_{k,t}$ admits a computable certificate $\mathcal{G}_{k,t}$ defined by (15), then*

$$\left\| \frac{1}{\gamma}(x_k - y_{k,t}) - \nabla \mathcal{U}_\gamma(x_k) \right\| \le \sqrt{\frac{2\mathcal{G}_{k,t}}{\gamma}}. \tag{16}$$

Consequently, the stopping rule

$$\mathcal{G}_{k,t} \le \frac{\gamma}{2}\varepsilon_k^2, \tag{17}$$

guarantees that the induced Moreau-gradient error is at most ε_k, i.e., $\delta_k \le \varepsilon_k$.

The certificate (15) is particularly convenient for primal–dual inner solvers. If $D_k(q_{k,t})$ is a valid dual objective value satisfying $D_k(q_{k,t}) \le F_k(y_k^\star)$, then one may simply take $\ell_{k,t} = D_k(q_{k,t})$ and $\mathcal{G}_{k,t} = F_k(y_{k,t}) - D_k(q_{k,t})$, which is the standard primal–dual gap. Whenever the inner method generates $\mathcal{G}_{k,t} \to 0$, the stopping rule (17) is attained after finite inner iterations for $\varepsilon_k > 0$.

Condition 3. Let $\mathcal{U} : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lower semicontinuous, and convex, with bounded subgradients: there exists $L > 0$ such that $|v| \le L$ for all $v \in \partial\mathcal{U}(x)$ and $x \in \mathbb{R}^n$. For a fixed outer iterate x_k, consider the proximal objective F_k in (14), whose unique minimizer is $y_k^\star = \mathbf{Prox}_{\gamma\mathcal{U}}(x_k)$. Suppose that the inner sequence $y_{k,t} * t \ge 0$ is generated by the subgradient method in Appendix D. Under the assumptions of Theorem D.1, $y * k, t \to y_k^\star$ as $t \to \infty$. Consequently, the induced Moreau-gradient error

$$\delta_{k,t} := \left\| \frac{1}{\gamma}(x_k - y_{k,t}) - \nabla \mathcal{U}_\gamma(x_k) \right\| = \frac{1}{\gamma}\|y_{k,t} - y_k^\star\|$$

satisfies $\delta_{k,t} \to 0$. Hence, for every prescribed $\varepsilon_k > 0$, there exists T_k such that $\delta_{k,t} \le \varepsilon_k$ for all $t \ge T_k$. Since $\delta_{k,t}$ depends on the unknown $y_k^\star$, this asymptotic result does not provide a computable finite-time certificate. In practice, quantities such as $|y_{k,t+1} - y_{k,t}|$ can serve as practical heuristic indicators of inner-iteration stabilization, as they typically decrease as the iterates approach $y_k^\star$.

3 Inexact proximal unadjusted Langevin algorithm

Markov chain Monte Carlo (MCMC) methods enable sampling from high-dimensional, complex distributions by constructing a Markov chain whose stationary distribution is the target density, $\pi(x)$. Under standard regularity conditions, the distribution of the generated samples converges to

this stationary regime as the chain evolves. To improve efficiency and convergence speed, MCMC algorithms can be augmented with gradient-based information; however, the composite potential function $\mathcal{U}$ of (6) is usually nonsmooth. As such, we first apply a smoothing technique based on the Moreau envelope (1) and then replace $\mathcal{U}$ with $\mathcal{U}_\gamma$ in (6), leading to $\pi_\gamma(x) := \frac{e^{-\mathcal{U}_\gamma(x)}}{\int_{\mathbb{R}^n} e^{-\mathcal{U}_\gamma(y)}\,dy}$. The (overdamped) Langevin dynamics for target distribution $\pi_\gamma(x)$ is a continuous-time stochastic process $(x_t)_{t\geq 0}$ in $\mathbb{R}^n$ that evolves the stochastic differential equation

$$dx_t = -\nabla \mathcal{U}_\gamma(x_t)dt + \sqrt{2}\, d\mathcal{B}_t, \tag{18}$$

where $\{\mathcal{B}_t\}_{t\geq 0}$ is a n-dimensional Brownian motion. It has $\pi_\gamma(x)$ as its stationary distribution, and its formulation allows sampling methods to use discretization and gradient-based techniques. Applying the Euler-Maruyama discretization scheme to the stochastic differential equation (SDE) (18) leads to the discrete-time Markov chain $\{x_k\}$ given by

$$x_{k+1} = x_k - \eta \nabla \mathcal{U}_\gamma(x_k) + \sqrt{2\eta}\mathcal{Z}_k, \quad k \in \mathbb{N}_0, \tag{19}$$

where $\eta > 0$ is the step-size, and $\{\mathcal{Z}_k\}_{k\geq 0}$ is a sequence of independent and identically distributed (i.i.d.) standard n-dimensional Gaussian random variables. Considering the composite form of the potential function $\mathcal{U}$, the proximal term (3) can only be computed inexactly, which consequently leads to the inexact gradient (9). As a result, (19) can be replaced by

$$x_{k+1} = x_k - \eta \nabla \mathcal{U}_\gamma^{\varepsilon_k}(x_k) + \sqrt{2\eta}\mathcal{Z}_k, \quad k \in \mathbb{N}_0, \tag{20}$$

which leads to the next inexact proximal-based unadjusted Langevin algorithm.

Algorithm 1 iPULA (Inexact Proximal Unadjusted Langevin Algorithm)

1: **Initialization** initial point $x_0 \in \mathbb{R}^n$, $\gamma > 0$, $\eta \in (0, 1/L_\gamma)$, tolerance sequence $\{\varepsilon_k\}_{k\in\mathbb{N}_0}$;
2: **while** stopping criteria do not hold **do**
3: Solve the subproblem approximately and return $y_k^{\varepsilon_k}$ such that $\delta_k \leq \varepsilon_k$;
4: Draw $\mathcal{Z}_k \sim \mathcal{N}(0, I_n)$ independently;
5: Set $x_{k+1} = \left(1 - \frac{\eta}{\gamma}\right) x_k + \frac{\eta}{\gamma} y_k^{\varepsilon_k} + \sqrt{2\eta}\mathcal{Z}_k$;
6: $k = k + 1$;
7: **end while**

For the stochastic analysis, let $\mathcal{F}_k$ denote the σ-algebra generated by the chain history up to iteration k. The oracle error e_k may depend on the chain history and the oracle's internal randomness, while the Gaussian innovation $\mathcal{Z}_k$ is drawn independently afterward.

3.1 Convergence analysis under convexity

Let $\mu_k := \mathcal{L}(x_k)$ denote the probability law of the kth iPULA iterate and define the averaged law

$$\nu_N := \frac{1}{N}\sum_{k=1}^{N} \mu_k.$$

We set $W_k := \mathcal{W}_2(\mu_k, \pi_\gamma)$ and assume that the possibly biased and history-dependent oracle errors satisfy $\mathbb{E}\,\|e_k\|^2 \leq \tau_k^2$ for deterministic $\tau_k \geq 0$. We next study the convex free-energy analysis of ULA in the Wasserstein space, extending Durmus et al. (2019) to the inexact setting by allowing biased, potentially history-dependent oracle errors.

Theorem 3.1 (**General-convex biased-iPULA bound**). *Suppose Assumption 2.1 holds, $\mathcal{L}(x_0) \in \mathcal{P}_2(\mathbb{R}^n)$, $0 < \eta < L_\gamma^{-1}$, and $\mathbb{E}\,\|e_k\|^2 \leq \tau_k^2$ for deterministic $\tau_k \geq 0$. Then*

$$D_{\mathrm{KL}}\left(\nu_N \| \pi_\gamma\right) \leq \frac{W_0^2}{2\eta N} + L_\gamma n\,\eta + \frac{1}{N}\sum_{k=0}^{N-1} \tau_k \sqrt{W_0^2 + 2L_\gamma n\,\eta^2 k} + \frac{\eta}{2N}\left(\sum_{k=0}^{N-1}\tau_k\right)^2 + \frac{L_\gamma \eta^2}{2(1 - L_\gamma \eta)N}\sum_{k=0}^{N-1}\tau_k^2. \tag{21}$$

Corollary 3.2 (**Square-summable inexactness**). *Under the assumptions of Theorem 3.1, if additionally* $\sum_{k=0}^{\infty} \tau_k^2 < \infty$, *then*

$$\limsup_{N\to\infty} D_{\mathrm{KL}}(\nu_N \| \pi_\gamma) \le L_\gamma n\, \eta. \tag{22}$$

Moreover, $\mathrm{TV}(\nu_N, \pi_\gamma) \le \sqrt{D_{\mathrm{KL}}(\nu_N\|\pi_\gamma)/2}$, *and*

$$\mathrm{TV}(\nu_N, \pi) \le \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(\nu_N\|\pi_\gamma)} + \mathrm{TV}(\pi_\gamma, \pi). \tag{23}$$

3.2 Convergence analysis under Polyak–Łojasiewicz inequality

Let us assume that the Moreau envelope satisfies the global Polyak–Łojasiewicz (PL) inequality

$$\frac{1}{2}\,\|\nabla \mathcal{U}_\gamma(x)\|^2 \ge \mu_\gamma\big(\mathcal{U}_\gamma(x) - \mathcal{U}_\gamma^\star\big), \quad \forall x \in \mathbb{R}^n, \tag{24}$$

for some $\mu_\gamma > 0$. A sufficient condition on the original potential is the nonsmooth PL inequality $\mathbf{dist}(0, \partial \mathcal{U}(x))^2 \ge 2\mu(\mathcal{U}(x) - \mathcal{U}^\star)$. For convex $\mathcal{U}$, the known Moreau-envelope inheritance result gives $\mu_\gamma = \mu/(1+\gamma\mu)$ (Apidopoulos et al., 2022). We also define

$$\rho_{\mathrm{PL}} := 1 - \mu_\gamma \eta (1 - 2L_\gamma \eta). \tag{25}$$

Theorem 3.3 (**Fixed-tolerance objective-gap bound**). *Suppose Assumption 2.1 and the PL inequality (24) hold. Let* $\{x_k\}_{k\in\mathbb{N}_0}$ *be generated by the inexact chain (20) with* $\eta \in (0, \frac{1}{2L_\gamma})$. *Assume further that there exists* $\delta \ge 0$ *such that* $\mathbb{E}\left[\delta_k^2 \,\middle|\, \mathcal{F}_k\right] \le \delta^2$, *for all* $k \in \mathbb{N}_0$. *Then, for any* $x^\star \in \mathbf{arg\,min}\, \mathcal{U}_\gamma$,

$$\mathbb{E}\left[\mathcal{U}_\gamma(x_{k+1}) - \mathcal{U}_\gamma(x^\star)\right] \le \rho_{\mathrm{PL}}\, \mathbb{E}\left[\mathcal{U}_\gamma(x_k) - \mathcal{U}_\gamma(x^\star)\right] + \eta\Big(\frac{1}{2} + L_\gamma \eta\Big)\delta^2 + \eta L_\gamma n, \tag{26}$$

and, for every $k \in \mathbb{N}_0$,

$$\mathbb{E}\left[\mathcal{U}_\gamma(x_k) - \mathcal{U}_\gamma(x^\star)\right] \le \rho_{\mathrm{PL}}^k \mathbb{E}\left[\mathcal{U}_\gamma(x_0) - \mathcal{U}_\gamma(x^\star)\right] + \frac{\eta(\frac{1}{2} + L_\gamma\eta)\delta^2 + \eta L_\gamma n}{1 - \rho_{\mathrm{PL}}}. \tag{27}$$

Theorem 3.4 (**Adaptive-tolerance objective-gap bound**). *Suppose Assumption 2.1 and the PL inequality (24) hold. Let* $\{x_k\}_{k\in\mathbb{N}_0}$ *be generated by the inexact chain (20) with* $\eta \in (0, \frac{1}{2L_\gamma})$. *Assume further that there exists a deterministic sequence* $\{\tau_k\}_{k\in\mathbb{N}_0}$ *such that* $\mathbb{E}\left[\delta_k^2 \,\middle|\, \mathcal{F}_k\right] \le \tau_k^2$, *for all* $k \in \mathbb{N}_0$. *Then, for any* $x^\star \in \mathbf{arg\,min}\, \mathcal{U}_\gamma$ *and for every* $k \in \mathbb{N}_0$,

$$\mathbb{E}\left[\mathcal{U}_\gamma(x_k) - \mathcal{U}_\gamma(x^\star)\right] \le \rho_{\mathrm{PL}}^k \mathbb{E}\left[\mathcal{U}_\gamma(x_0) - \mathcal{U}_\gamma(x^\star)\right] + \eta\left(\frac{1}{2} + L_\gamma\eta\right)\sum_{j=0}^{k-1} \rho_{\mathrm{PL}}^{k-1-j}\tau_j^2 + \eta L_\gamma n \frac{1-\rho_{\mathrm{PL}}^k}{1-\rho_{\mathrm{PL}}}. \tag{28}$$

3.3 Convergence analysis under strong convexity

For the remainder of this subsection, suppose additionally that $\mathcal{U}$ is σ-strongly convex. By Fact 2.3, $\mathcal{U}_\gamma$ is m_γ-strongly convex with modulus $m_\gamma := \frac{\sigma}{1+\gamma\sigma}$. Let us define

$$q := \sqrt{1 - 2m_\gamma\eta + L_\gamma^2\eta^2}, \quad 0 < \eta < \mathbf{min}\left\{L_\gamma^{-1}, \frac{2m_\gamma}{L_\gamma^2}, \frac{2}{m_\gamma + L_\gamma}\right\}. \tag{29}$$

implying $q \in [0, 1)$. Applying the standard constant-step exact-LMC estimate of Dalalyan & Karagulyan (2019, Theorem 1(a)) to $\mathcal{U}_\gamma$ gives

$$\mathcal{W}_2(\mathcal{L}(\overline{x}_k), \pi_\gamma) \le (1 - m_\gamma\eta)^k \mathcal{W}_2(\mathcal{L}(x_0), \pi_\gamma) + 1.65\frac{L_\gamma}{m_\gamma}\sqrt{\eta n}. \tag{30}$$

Combining this estimate with Theorem J.2 gives the following closed iPULA result.

Theorem 3.5 (Inexact Wasserstein convergence). *Let the assumptions of Theorem J.2 hold. Then, for every* $k \in \mathbb{N}_0$,

$$\mathcal{W}_2\left(\mathcal{L}(x_k), \pi_\gamma\right) \le (1 - m_\gamma\eta)^k \mathcal{W}_2\left(\mathcal{L}(x_0), \pi_\gamma\right) + 1.65\frac{L_\gamma}{m_\gamma}\sqrt{\eta n} + \eta\sum_{j=0}^{k-1} q^{k-1-j}\tau_j, \tag{31}$$

where q *is given in (29). In particular, for the fixed schedule* $\tau_j \equiv \delta$, *this yields*

$$\mathcal{W}_2(\mathcal{L}(x_k), \pi_\gamma) \le (1 - m_\gamma\eta)^k \mathcal{W}_2(\mathcal{L}(x_0), \pi_\gamma) + 1.65\frac{L_\gamma}{m_\gamma}\sqrt{\eta n} + \frac{\eta\delta(1-q^k)}{1-q}. \tag{32}$$

4 EXPERIMENTS

We evaluate iPULA on two imaging inverse problems with nonsmooth total variation (TV) regularization where $\mathbf{Prox}_{\gamma\mathcal{U}}$ lacks a closed form: (i) Gaussian MRI deblurring, and (ii) Poisson deconvolution on real confocal microscopy from the FMD benchmark (Zhang et al., 2019). We compare iPULA against MYULA (Durmus et al., 2018) and splitting schemes Grad-sub and Prox-sub (Habring et al., 2024), using inner projected subgradient iterations (Mordukhovich & Nam, 2014).

4.1 EXPERIMENTAL SETUP AND PROBLEM FORMULATIONS

Application 1: Brain MRI Deblurring under Gaussian Noise. Given ground-truth $x_{\text{true}} \in \mathbb{R}^{n\times n}$ (128×128), degraded observations are $y = Hx_{\text{true}} + \varepsilon$, where H models a 5×5 blur and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ is additive Gaussian noise (blurred SNR 40 dB). We sample from the posterior potential

$$U_{\text{Gauss}}(x) = \frac{1}{2\sigma^2}\|Hx - y\|_2^2 + \lambda_{\text{TV}}\text{TV}(x), \quad \text{with } \lambda_{\text{TV}} = 10^{-3}. \tag{33}$$

All methods run for 5000 iterations with $\eta = 0.4 \times 10^{-6}$ and $\gamma = 10^{-6}$ (500 burn-in for posterior mean estimation, and the stationary second half of 2500 draws for mixing analysis).

Application 2: Real Fluorescence Microscopy under Poisson Noise. Confocal images of bovine pulmonary artery endothelial (BPAE) cells (128×128) are degraded by a 5×5 optical blur operator H and acquired under Poisson counting statistics: $y_i \sim \text{Poisson}([I_{\text{peak}}Hx_{\text{true}}]_i + b)$ with $I_{\text{peak}} = 1000$ and $b = 1.0$. The data fidelity is given by the Poisson deviance $f(x) := D_{\text{KL}}(y \parallel \mu(x)) = \sum_{i=1}^{n^2}[y_i \log(y_i/\mu_i(x)) - (y_i - \mu_i(x))]$ with $\mu(x) := I_{\text{peak}}Hx + b > 0$. The composite potential is

$$U_{\text{Pois}}(x) = D_{\text{KL}}(y \parallel \mu(x)) + \lambda_{\text{TV}}\text{TV}(x), \quad \text{with } \lambda_{\text{TV}} = 10^{-2}. \tag{34}$$

Chains run for 1000 iterations with $\gamma = 10^{-6}$ and $\eta = 0.4 \times 10^{-6}$ (250 burn-in for posterior mean estimation, and the stationary second half of 500 draws for mixing analysis).

Table 1: Sampling efficiency and reconstruction quality over 5000 BrainWeb MRI and 1000 BPAE fluorescence microscopy iterations. Metrics include peak reconstruction, posterior mean PSNR, time to peak, IACT, and ESS, evaluated after 50% burn-in (2500 MRI and 500 microscopy draws).

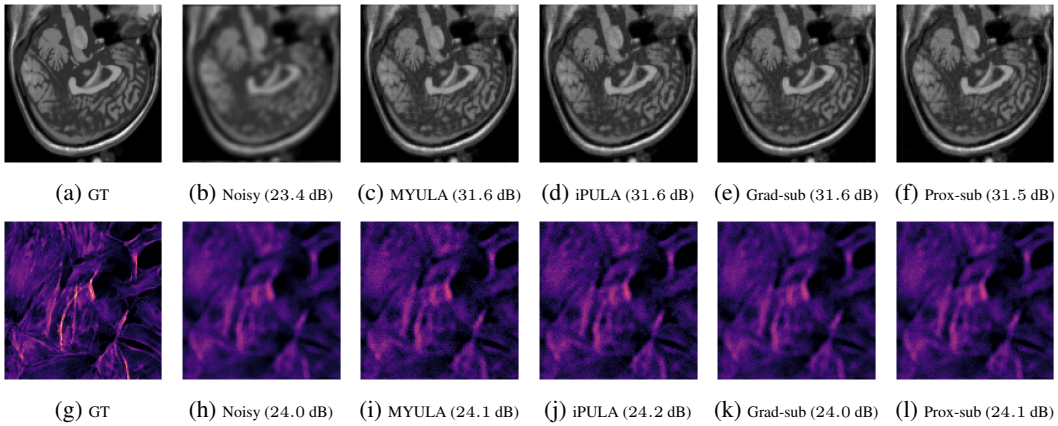
Method	Best PSNR (dB)	Best SSIM	Mean PSNR (dB)	Time to Best (s)	IACT	ESS
Application 1: Brain MRI Deblurring under Gaussian Noise						
Blurred-noisy	23.42	0.7545	–	–	–	–
iPULA (Ours) m	**31.57**	**0.9361**	31.42	0.060	**42.85**	**58**
Grad-sub	31.57	0.9342	31.41	0.110	54.72	46
Prox-sub	31.54	0.9349	31.39	0.103	48.36	52
MYULA	31.65	0.9429	31.50	0.020	48.24	52
Application 2: BPAE Fluorescence Microscopy under Poisson Noise						
Poisson-noisy	23.96	0.6030	–	–	–	–
iPULA (Ours)	**24.21**	**0.6205**	**24.31**	4.714	**53.31**	**9**
Grad-sub	24.04	0.6110	24.08	2.286	85.28	6
Prox-sub	24.07	0.6121	24.10	5.540	84.18	6
MYULA	24.12	0.6124	24.25	2.868	67.79	7

4.2 EMPIRICAL ANALYSIS AND MIXING DYNAMICS

Reconstruction Fidelity and Efficiency.

Table 1 and Figure 1 summarize quantitative and visual performance across both benchmarks. For Gaussian MRI deblurring, all methods restore fine anatomical boundaries from the degraded input (23.42 dB, SSIM 0.7545), with iPULA achieving 31.57 dB (SSIM 0.9361) and its peak estimate in 0.060 s (iteration 5). In Poisson microscopy, the raw image is heavily degraded by noise (23.96 dB, SSIM 0.6030). iPULA achieves a peak PSNR of 24.21 dB (SSIM 0.6205) and posterior mean PSNR of 24.31 dB, compared to MYULA (24.12 dB), Prox-sub (24.07 dB), and Grad-sub (24.04 dB).

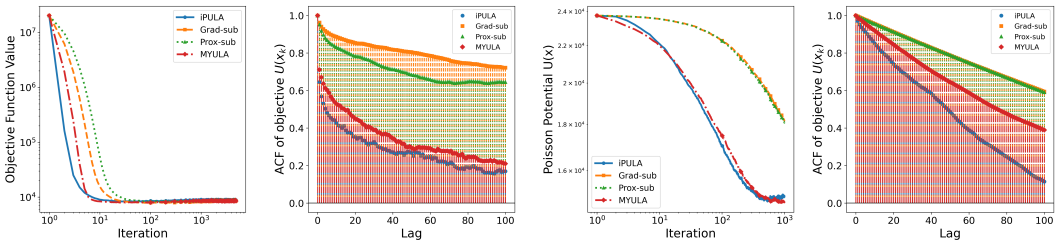
Autocorrelation and Posterior Exploration. Figure 2 shows convergence profiles and autocorrelation decay of the stationary chains. On the Gaussian problem, iPULA achieves rapid ACF decay (to 0.17 at lag 100), with IACT $= 42.85$ and ESS $= 58$, compared to 54.72 (ESS $= 46$) for Grad-sub.



(a) GT (b) Noisy (23.4 dB) (c) MYULA (31.6 dB) (d) iPULA (31.6 dB) (e) Grad-sub (31.6 dB) (f) Prox-sub (31.5 dB)

(g) GT (h) Noisy (24.0 dB) (i) MYULA (24.1 dB) (j) iPULA (24.2 dB) (k) Grad-sub (24.0 dB) (l) Prox-sub (24.1 dB)

Figure 1: Visual reconstructions. Top row: BrainWeb MRI deblurring under Gaussian noise. Bottom row: BPAE cell deconvolution under Poisson noise.



(a) BrainWeb $U(x)$ (b) BrainWeb ACF (c) BPAE deviance (d) BPAE ACF

Figure 2: Potential value $U(x)$ versus iteration and autocorrelation function (ACF) of the post-burn-in stationary chain for BrainWeb MRI (a, b) and BPAE microscopy (c, d).

This performance differential is amplified under Poisson counting noise (Figure 2d). Because the Poisson deviance gradient $\nabla f(x) = I_{\text{peak}} H^\top (\mathbf{1} - y/\mu(x))$ is non-linear and non-Lipschitz near zero, splitting algorithms such as Grad-sub and Prox-sub decouple the likelihood gradient from the total variation prior, leading to autocorrelation stagnation (ACF ≈ 0.60 at lag 100, IACT ≈ 85). In contrast, iPULA computes an approximate proximal step over the joint composite objective $U(x) = f(x) + g(x)$, dampening likelihood oscillations, reducing lag-100 autocorrelation to 0.11, and achieving the lowest IACT (53.31) and highest effective sample size (ESS $= 9$).

5 Discussion

In this work, we considered sampling from posterior distributions with nonsmooth composite potentials whose nonsmooth component lacks a closed-form proximal operator. We approximated the proximal operator to construct an inexact Moreau-gradient oracle and proposed iPULA, an inexact proximal unadjusted Langevin algorithm. Under suitable proximal inexactness criteria, we established non-asymptotic convergence guarantees, while experiments on image recovery with total variation regularization demonstrated strong practical performance.

Our framework is flexible and only requires the proximal subproblem to be solved efficiently. In the convex setting, various methods can be used, including subgradient schemes Davis et al. (2018); Mordukhovich & Nam (2014); Rahimi et al. (2026); however, identifying the most efficient approach in practice remains an open question.

AI use statement

In this work, we used generative AI tools for language editing.

REFERENCES

Masoud Ahookhosh, Andreas Themelis, and Panagiotis Patrinos. A Bregman forward-backward line-search algorithm for nonconvex composite optimization: Superlinear convergence to nonisolated local minima. *SIAM Journal on Optimization*, 31(1):653–685, 2021.

Vassilis Apidopoulos, Nicolò Ginatta, and Silvia Villa. Convergence rates for the heavy-ball continuous dynamics for non-convex optimization, under Polyak–Łojasiewicz condition. *Journal of Global Optimization*, 84:563–589, 2022.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2nd edition, 2017.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng (eds.). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

Niladri S. Chatterji, Jelena Diakonikolas, Michael I. Jordan, and Peter L. Bartlett. Langevin monte carlo without smoothness. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1716–1726, 2020.

F. R. Crucinio, Alain Durmus, Pablo Jiménez, and Gareth O. Roberts. Optimal scaling results for Moreau–Yosida metropolis-adjusted Langevin algorithms. *Bernoulli*, 31(3):1889–1907, 2025.

Arnak S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 678–689, 2017.

Arnak S. Dalalyan and Avetik G. Karagulyan. User-friendly guarantees for the Langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

Arnak S. Dalalyan, Avetik Karagulyan, and Lionel Riou-Durand. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *Journal of Machine Learning Research*, 23(235):1–38, 2022.

Damek Davis, Dmitriy Drusvyatskiy, Kellie J. MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3): 962–982, 2018.

Kumar Avinava Dubey, Sashank J. Reddi, Sinead A. Williamson, Barnabás Póczos, Alexander J. Smola, and Eric P. Xing. Variance reduction in stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Alain Durmus, Éric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

Armin Eftekhari, Luis Vargas, and Konstantinos C. Zygalakis. The forward-backward envelope for sampling with the overdamped Langevin algorithm. *Statistics and Computing*, 33(4):85, 2023.

Matthias J. Ehrhardt, Lorenz Kuger, and Carola-Bibiane Schönlieb. Proximal Langevin sampling with inexact proximal mapping. *SIAM Journal on Imaging Sciences*, 17(3):1729–1760, 2024.

Susan Ghaderi, Masoud Ahookhosh, Adam Arany, Alexander Skupin, Panagiotis Patrinos, and Yves Moreau. Smoothing unadjusted Langevin algorithms for nonsmooth composite potential functions. *Applied Mathematics and Computation*, 464:128377, 2024.

Andreas Habring, Martin Holler, and Thomas Pock. Subgradient Langevin methods for sampling from nonsmooth potentials. *SIAM Journal on Mathematics of Data Science*, 6(4):897–925, 2024.

Jonathan Huggins and James Zou. Quantifying the accuracy of approximate diffusions and Markov chains. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 382–391. PMLR, 2017.

Alireza Kabgani and Masoud Ahookhosh. ItsDEAL: Inexact two-level smoothing descent algorithms for weakly convex optimization. *arXiv: 2501.02155*, 2025a.

Alireza Kabgani and Masoud Ahookhosh. Moreau envelope and proximal-point methods under the lens of high-order regularization. *Set-Valued and Variational Analysis*, 33:47, 2025b.

Alireza Kabgani and Masoud Ahookhosh. ItsOPT: An inexact two-level smoothing framework for nonconvex optimization via high-order Moreau envelope. *SIAM Journal on Optimization (to appear)*, 2026a.

Alireza Kabgani and Masoud Ahookhosh. On fundamental properties of high-order forward-backward envelope. *Journal of Optimization Theory and Applications*, 210(1):14, 2026b.

Teresa Klatzer, Paul Dobson, Yoann Altmann, Marcelo Pereyra, Jesús María Sanz-Serna, and Konstantinos C. Zygalakis. Accelerated Bayesian imaging by relaxed proximal-point Langevin sampling. *SIAM Journal on Imaging Sciences*, 17(2):1078–1117, 2024.

Tim Tsz-Kit Lau and Han Liu. Bregman proximal Langevin monte carlo via Bregman–Moreau envelopes. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12049–12077, 2022.

Tim Tsz-Kit Lau, Han Liu, and Thomas Pock. Non-log-concave and nonsmooth sampling via Langevin monte carlo algorithms. In *Advanced Techniques in Optimization for Machine Learning and Imaging*, volume 61 of *Springer INdAM Series*, pp. 83–149. Springer, 2024.

Yi-An Ma, Tianqi Chen, and Emily B. Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Boris S. Mordukhovich and Nguyen Mau Nam. *An Easy Path to Convex Analysis and Applications*. Morgan & Claypool, Williston, 2014.

Jean-Jacques Moreau. Propriétés des applications "prox". *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 256:1069–1071, 1963.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Marcelo Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4): 745–760, 2016.

Chayne Planiden and Xianfu Wang. Strongly convex functions, Moreau envelopes, and the generic nature of convex functions with strong minimizers. *SIAM Journal on Optimization*, 26(2):1341–1364, 2016.

Morteza Rahimi, Susan Ghaderi, Yves Moreau, and Masoud Ahookhosh. Projected subgradient methods for paraconvex optimization: Application to robust low-rank matrix recovery. *Journal of Global Optimization*, pp. 1–40, 2026.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2004.

Adil Salim, Dmitry Kovalev, and Peter Richtárik. Stochastic proximal Langevin algorithm: Potential splitting and nonasymptotic rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Apratim Shukla, Dootika Vats, and Eric C. Chi. MCMC importance sampling via Moreau–Yosida envelopes, 2025.

Lorenzo Stella, Andreas Themelis, and Panagiotis Patrinos. Forward-backward quasi-Newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 67 (3):443–487, 2017.

Andreas Themelis, Masoud Ahookhosh, and Panagiotis Patrinos. On the acceleration of forward-backward splitting via an inexact Newton method. In *Splitting Algorithms, Modern Operator Theory, and Applications*, pp. 363–412. Springer, 2019.

Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A poisson–gaussian denoising dataset with real fluorescence microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11710–11718, 2019.

CONTENTS

1 Introduction **1**

1.1 Related work: Nonsmooth MCMC methods and key question 2

1.2 Contribution 3

1.3 Notation 3

2 Problem setup and Moreau envelope **3**

2.1 Proximal operator and Moreau envelope 4

2.2 Inexact first-order oracle for Moreau envelope 4

2.3 Sufficient certificate for acceptable inexact proximal term 4

3 Inexact proximal unadjusted Langevin algorithm **5**

3.1 Convergence analysis under convexity 6

3.2 Convergence analysis under Polyak–Łojasiewicz inequality 7

3.3 Convergence analysis under strong convexity 7

4 Experiments **8**

4.1 Experimental Setup and Problem Formulations 8

4.2 Empirical Analysis and Mixing Dynamics 8

5 Discussion **9**

A Additional preliminaries **15**

B Proof of Proposition 2.4 **15**

C Proof of Proposition 2.5 **16**

D Proximal certificate via subgradient method **16**

E Lemma E.1 **18**

F Proof of Theorem 3.1 **19**

G Proof of Corollary 3.2 **19**

H Proof of Theorem 3.3 **20**

I Proof of Theorem 3.4 **22**

J Inexact iPULA chain vs. exact Moreau–ULA chain **22**

K Proof of Theorem 3.5 **24**

L Consequences of adaptive tolerance schedules **24**

A Additional preliminaries

Definition A.1. *A function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ with convex domain is said to be strongly convex on $\mathbf{dom}(h)$ if there exists $\sigma > 0$ such that, for all $x, y \in \mathbf{dom}(h)$ and all $\lambda \in [0, 1]$,*

$$h(\lambda y + (1-\lambda)x) \leq \lambda h(y) + (1-\lambda)h(x) - \lambda(1-\lambda)\frac{\sigma}{2}\|x-y\|^2, \tag{35}$$

For a proper convex function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$, the convex subdifferential of h at $x \in \mathbb{R}^n$ is defined by

$$\partial h(x) := \{v \in \mathbb{R}^n : h(y) \geq h(x) + \langle v, y - x\rangle \quad \forall y \in \mathbb{R}^n\},$$

with the convention $\partial h(x) = \emptyset$ if $x \notin \mathbf{dom}(h)$. Any vector $v \in \partial h(x)$ is called a subgradient of h at x. The mapping ∂h is monotone, that is,

$$\langle u - v, x - y\rangle \geq 0, \quad \forall u \in \partial h(x),\ v \in \partial h(y).$$

Fact 2.2 records the differentiability properties of the Moreau envelope. We next recall additional properties that will be used throughout the analysis.

Fact A.2 (Further properties of the Moreau envelope). *(Bauschke & Combettes, 2017) Suppose Assumption 2.1 holds and let $\gamma > 0$. Then,*

- **(a)** $\mathbf{dom}(\mathcal{U}_\gamma) = \mathbb{R}^n$ *and $\mathcal{U}_\gamma$ is finite-valued;*
- **(b)** *for every $x \in \mathbb{R}^n$, $\mathbf{inf}_{y\in\mathbb{R}^n}\, \mathcal{U}(y) \leq \mathcal{U}_\gamma(x) \leq \mathcal{U}(x)$;*
- **(c)** $\mathbf{inf}_{x\in\mathbb{R}^n}\, \mathcal{U}_\gamma(x) = \mathbf{inf}_{x\in\mathbb{R}^n}\, \mathcal{U}(x)$*;*
- **(d)** $\mathbf{Prox}_{\gamma\mathcal{U}}(x)$ *is nonempty and single-valued for every $x \in \mathbb{R}^n$;*
- **(e)** *$\mathcal{U}_\gamma$ is convex, real-valued, and continuous;*

B Proof of Proposition 2.4

Proof. Fix $x_k \in \mathbb{R}^n$ and let $y_k := \mathbf{Prox}_{\gamma\mathcal{U}}(x_k)$ be the exact proximal point. Since $\mathcal{U}$ is proper, lower semicontinuous, and convex, the proximal subproblem is strongly convex and therefore admits a unique minimizer. Its first-order optimality condition gives

$$0 \in \partial\mathcal{U}(y_k) + \frac{1}{\gamma}(y_k - x_k).$$

Hence, there exists $v_k \in \partial\mathcal{U}(y_k)$ such that

$$v_k + \frac{1}{\gamma}(y_k - x_k) = 0. \tag{36}$$

By the definition of the residual r_k, and since $y_k^{\varepsilon_k} \in \mathbf{dom}(\partial\mathcal{U})$, there exists $u_k \in \partial\mathcal{U}(y_k^{\varepsilon_k})$ such that

$$\zeta_k := u_k + \frac{1}{\gamma}(y_k^{\varepsilon_k} - x_k), \quad \|\zeta_k\| = r_k. \tag{37}$$

If the distance in (11) is not attained, the same argument applies with $\|\zeta_k\| \leq r_k + \rho$ for arbitrary $\rho > 0$, followed by letting $\rho \downarrow 0$.

Subtracting (36) from (37) yields

$$u_k - v_k + \frac{1}{\gamma}(y_k^{\varepsilon_k} - y_k) = \zeta_k.$$

Taking the inner product with $y_k^{\varepsilon_k} - y_k$ gives

$$\langle u_k - v_k, y_k^{\varepsilon_k} - y_k\rangle + \frac{1}{\gamma}\|y_k^{\varepsilon_k} - y_k\|^2 = \langle \zeta_k, y_k^{\varepsilon_k} - y_k\rangle.$$

Since $\mathcal{U}$ is convex, its subdifferential $\partial\mathcal{U}$ is monotone. Therefore,

$$\langle u_k - v_k, y_k^{\varepsilon_k} - y_k \rangle \geq 0.$$

It follows that

$$\frac{1}{\gamma} \left\| y_k^{\varepsilon_k} - y_k \right\|^2 \leq \langle \zeta_k, y_k^{\varepsilon_k} - y_k \rangle \leq \|\zeta_k\| \, \|y_k^{\varepsilon_k} - y_k\| .$$

If $y_k^{\varepsilon_k} = y_k$, the desired inequality is immediate. Otherwise, dividing by $\|y_k^{\varepsilon_k} - y_k\|$ gives

$$\frac{1}{\gamma} \|y_k^{\varepsilon_k} - y_k\| \leq \|\zeta_k\| = r_k.$$

Finally, by the definitions of the exact and inexact Moreau-gradient oracles,

$$\nabla \mathcal{U}_\gamma^{\varepsilon_k}(x_k) - \nabla \mathcal{U}_\gamma(x_k) = \frac{1}{\gamma} \left(y_k - y_k^{\varepsilon_k} \right).$$

Thus,

$$\delta_k = \left\| \nabla \mathcal{U}_\gamma^{\varepsilon_k}(x_k) - \nabla \mathcal{U}_\gamma(x_k) \right\| = \frac{1}{\gamma} \|y_k^{\varepsilon_k} - y_k\| \leq r_k.$$

Finally, $r_k \leq \widehat{r}_k$ follows immediately from the definition of the distance in (11) by evaluating it at the available selection $u_k \in \partial \mathcal{U}(y_k^{\varepsilon_k})$, which adjusts (13). □

C Proof of Proposition 2.5

Proof. Since F_k is $1/\gamma$-strongly convex and $y_k^\star$ is its minimizer,

$$F_k(y_{k,t}) - F_k(y_k^\star) \geq \frac{1}{2\gamma} \|y_{k,t} - y_k^\star\|^2.$$

Moreover, $\ell_{k,t} \leq F_k(y_k^\star)$ implies

$$F_k(y_{k,t}) - F_k(y_k^\star) \leq F_k(y_{k,t}) - \ell_{k,t} = \mathcal{G}_{k,t}.$$

Therefore,

$$\|y_{k,t} - y_k^\star\| \leq \sqrt{2\gamma \mathcal{G}_{k,t}}.$$

Using

$$\nabla U_\gamma(x_k) = \frac{1}{\gamma}(x_k - y_k^\star),$$

we obtain

$$\left\| \frac{1}{\gamma}(x_k - y_{k,t}) - \nabla U_\gamma(x_k) \right\| = \frac{1}{\gamma} \|y_{k,t} - y_k^\star\| \leq \sqrt{\frac{2\mathcal{G}_{k,t}}{\gamma}},$$

which proves (16). The stopping rule (17) then gives the claimed ε_k-accuracy. □

D Proximal certificate via subgradient method

Fix an outer iterate $x_k \in \mathbb{R}^n$ and consider the proximal objective F_k as defined in (14). Since $\mathcal{U}$ is proper, lower semicontinuous, and convex, F_k is $1/\gamma$-strongly convex and admits the unique minimizer $y_k^\star = \mathbf{Prox}_{\gamma\mathcal{U}}(x_k)$. Assume further that the subgradients of $\mathcal{U}$ are uniformly bounded: there exists $L > 0$ such that

$$\|u\| \leq L, \qquad \forall u \in \partial \mathcal{U}(y), \quad y \in \mathbb{R}^n.$$

Theorem D.1 (Convergence of the inner subgradient method)**.** *For a fixed outer iteration k, let $\{y_{k,t}\}_{t \geq 0}$ be generated by*

$$y_{k,t+1} = y_{k,t} - \alpha_{k,t} d_{k,t}, \qquad d_{k,t} := u_{k,t} + \frac{1}{\gamma}(y_{k,t} - x_k), \qquad u_{k,t} \in \partial \mathcal{U}(y_{k,t}). \tag{38}$$

Suppose that

$$0 < \alpha_{k,t} \leq \gamma, \qquad \sum_{t=0}^{\infty} \alpha_{k,t} = +\infty, \qquad \sum_{t=0}^{\infty} \alpha_{k,t}^2 < +\infty.$$

Then

$$y_{k,t} \longrightarrow y_k^\star \qquad \text{as } t \to \infty.$$

Proof. Fix the outer iteration k. By the optimality condition for the proximal subproblem, there exists $u_k^\star \in \partial \mathcal{U}(y_k^\star)$ such that

$$u_k^\star + \frac{1}{\gamma}(y_k^\star - x_k) = 0. \tag{39}$$

Using (38) and (39), we obtain

$$\begin{aligned} y_{k,t+1} - y_k^\star &= y_{k,t} - y_k^\star - \alpha_{k,t}\left(u_{k,t} + \frac{1}{\gamma}(y_{k,t} - x_k)\right) \\ &= \left(1 - \frac{\alpha_{k,t}}{\gamma}\right)(y_{k,t} - y_k^\star) - \alpha_{k,t}(u_{k,t} - u_k^\star). \end{aligned} \tag{40}$$

Since $\|u_{k,t}\| \le L$ and $\|u_k^\star\| \le L$, and $0 < \alpha_{k,t} \le \gamma$, it follows that

$$\|y_{k,t+1} - y_k^\star\| \le \left(1 - \frac{\alpha_{k,t}}{\gamma}\right)\|y_{k,t} - y_k^\star\| + 2L\alpha_{k,t}. \tag{41}$$

Define

$$R_k := \mathbf{max}\left\{\|y_{k,0} - y_k^\star\|, 2\gamma L\right\}.$$

If $\|y_{k,t} - y_k^\star\| \le R_k$, then

$$\|y_{k,t+1} - y_k^\star\| \le \left(1 - \frac{\alpha_{k,t}}{\gamma}\right)R_k + 2L\alpha_{k,t} = R_k - \frac{\alpha_{k,t}}{\gamma}(R_k - 2\gamma L) \le R_k.$$

Hence, by induction,

$$\|y_{k,t} - y_k^\star\| \le R_k, \qquad \forall t \ge 0. \tag{42}$$

Thus, $\{y_{k,t}\}_{t\ge 0}$ is bounded. Consequently, the selected subgradients of F_k are bounded along the inner iterations. Indeed,

$$d_{k,t} = u_{k,t} + \frac{1}{\gamma}(y_{k,t} - x_k) \in \partial F_k(y_{k,t}),$$

and, by (42),

$$\|d_{k,t}\| \le L + \frac{1}{\gamma}\|y_{k,t} - x_k\| \le L + \frac{1}{\gamma}\left(R_k + \|y_k^\star - x_k\|\right) =: G_k,$$

where $G_k < +\infty$ is independent of t. Since F_k is $1/\gamma$-strongly convex, $d_{k,t} \in \partial F_k(y_{k,t})$, and $0 \in \partial F_k(y_k^\star)$, the strong monotonicity of ∂F_k yields

$$\langle d_{k,t}, y_{k,t} - y_k^\star\rangle \ge \frac{1}{\gamma}\|y_{k,t} - y_k^\star\|^2. \tag{43}$$

Therefore, using (38),

$$\begin{aligned} \|y_{k,t+1} - y_k^\star\|^2 &= \|y_{k,t} - y_k^\star\|^2 - 2\alpha_{k,t}\langle d_{k,t}, y_{k,t} - y_k^\star\rangle + \alpha_{k,t}^2\|d_{k,t}\|^2 \\ &\le \|y_{k,t} - y_k^\star\|^2 - \frac{2\alpha_{k,t}}{\gamma}\|y_{k,t} - y_k^\star\|^2 + G_k^2\alpha_{k,t}^2. \end{aligned} \tag{44}$$

Set

$$a_t := \|y_{k,t} - y_k^\star\|^2, \qquad b_t := G_k^2\alpha_{k,t}^2, \qquad c_t := \frac{2\alpha_{k,t}}{\gamma}a_t.$$

Then (44) becomes $a_{t+1} \le a_t - c_t + b_t$. Since

$$\sum_{t=0}^{\infty} b_t = G_k^2 \sum_{t=0}^{\infty} \alpha_{k,t}^2 < +\infty,$$

define

$$B_t := \sum_{j=t}^{\infty} b_j.$$

Then

$$a_{t+1} + B_{t+1} \le a_t + B_t - c_t.$$

Hence, $\{a_t + B_t\}_{t\geq 0}$ is nonincreasing and bounded below by zero, so it converges. Since $B_t \to 0$, the sequence $\{a_t\}$ also converges, say $a_t \to \bar{a} \geq 0$. Moreover, summing the preceding inequality gives $\sum_{t=0}^{\infty} c_t < +\infty$, and therefore

$$\sum_{t=0}^{\infty} \alpha_{k,t} \|y_{k,t} - y_k^\star\|^2 < +\infty. \tag{45}$$

Suppose, for contradiction, that $\bar{a} > 0$. Then there exists $T \geq 0$ such that

$$a_t \geq \frac{\bar{a}}{2}, \qquad \forall t \geq T.$$

Consequently,

$$\sum_{t=T}^{\infty} \alpha_{k,t} a_t \geq \frac{\bar{a}}{2} \sum_{t=T}^{\infty} \alpha_{k,t} = +\infty,$$

which contradicts (45). Hence $\bar{a} = 0$, and therefore $y_{k,t} \to y_k^\star$. Finally, by the Moreau-envelope gradient formula,

$$\nabla \mathcal{U}_\gamma(x_k) = \frac{1}{\gamma}(x_k - y_k^\star).$$

Thus,

$$\delta_{k,t} = \left\| \frac{1}{\gamma}(x_k - y_{k,t}) - \nabla \mathcal{U}_\gamma(x_k) \right\| = \frac{1}{\gamma} \|y_{k,t} - y_k^\star\| \longrightarrow 0,$$

which completes the proof. □

E LEMMA E.1

Lemma E.1 (Biased free-energy recursion). *Suppose Assumption 2.1 holds,* $0 < \eta < L_\gamma^{-1}$, $\mathbb{E}\,\|e_k\|^2 \leq \tau_k^2$ *for some deterministic* $\tau_k \geq 0$*, and* $\mu_k \in \mathcal{P}_2(\mathbb{R}^n)$*. Then*

$$2\eta\, D_{\mathrm{KL}}(\mu_{k+1} \| \pi_\gamma) \leq W_k^2 - W_{k+1}^2 + 2L_\gamma n\, \eta^2 + 2\eta\tau_k W_k + \frac{\eta^2}{1 - L_\gamma \eta} \tau_k^2. \tag{46}$$

Proof. For a deterministic state x, error vector e, and inexact drift point $z = x - \eta(\nabla \mathcal{U}_\gamma(x) + e)$, convexity and L_γ-smoothness imply, for every y,

$$\begin{aligned} 2\eta(\mathcal{U}_\gamma(z) - \mathcal{U}_\gamma(y)) \leq\ & \|x - y\|^2 - \|z - y\|^2 - 2\eta \langle e, x - y \rangle \\ & + \eta^2 \big[-(1 - L_\gamma\eta) \|\nabla \mathcal{U}_\gamma(x)\|^2 + 2L_\gamma \eta \langle \nabla \mathcal{U}_\gamma(x), e \rangle + (1 + L_\gamma \eta) \|e\|^2 \big]. \end{aligned}$$

Since $L_\gamma \eta < 1$,

$$-(1 - L_\gamma\eta) \|\nabla \mathcal{U}_\gamma(x)\|^2 + 2L_\gamma \eta \langle \nabla \mathcal{U}_\gamma(x), e \rangle \leq \frac{(L_\gamma \eta)^2}{1 - L_\gamma \eta} \|e\|^2,$$

and consequently

$$2\eta(\mathcal{U}_\gamma(z) - \mathcal{U}_\gamma(y)) \leq \|x - y\|^2 - \|z - y\|^2 - 2\eta \langle e, x - y \rangle + \frac{\eta^2}{1 - L_\gamma \eta} \|e\|^2. \tag{47}$$

Take an optimal $\mathcal{W}_2$ coupling (X, Y) of μ_k and π_γ, disintegrate it conditionally on X, and combine it with the regular conditional law of the inner error. The Cauchy–Schwarz inequality implies

$$|\mathbb{E} \langle e_k, X - Y \rangle| \leq \tau_k W_k.$$

Integrating (47) gives the perturbed deterministic-step free-energy inequality. The Gaussian heat step is the same entropy-flow step used in the exact convex analysis of Durmus et al. (2019). It contributes $2L_\gamma n \eta^2$ and telescopes the intermediate Wasserstein term. Adding the energy and entropy inequalities yields (46). □

F Proof of Theorem 3.1

Proof. Summing (46) from $k = 0$ to $N - 1$ and using convexity of KL under averaging gives

$$D_{\mathrm{KL}}(\nu_N \| \pi_\gamma) \le \frac{W_0^2}{2\eta N} + L_\gamma n\eta + \frac{1}{N}\sum_{k=0}^{N-1} \tau_k W_k + \frac{\eta}{2(1 - L_\gamma \eta)N} \sum_{k=0}^{N-1} \tau_k^2. \tag{48}$$

To bound W_k, couple iPULA and the exact chain (19) synchronously. Since $\mathcal{U}_\gamma$ is convex and L_γ-smooth, cocoercivity implies that $I - \eta\nabla\mathcal{U}_\gamma$ is nonexpansive for $0 \le \eta \le 2/L_\gamma$, and hence

$$\mathcal{W}_2(\mu_k, \mathcal{L}(\overline{x}_k)) \le \eta \sum_{j=0}^{k-1} \tau_j.$$

For the exact chain, the convex ($m = 0$) specialization of the one-step free-energy inequality of Durmus et al. (2019), together with nonnegativity of KL, gives

$$\mathcal{W}_2^2(\mathcal{L}(\overline{x}_k), \pi_\gamma) \le W_0^2 + 2L_\gamma n\eta^2 k.$$

Thus

$$W_k \le \sqrt{W_0^2 + 2L_\gamma n\eta^2 k} + \eta \sum_{j<k} \tau_j.$$

Substituting this bound into (48) and using

$$\sum_{k=0}^{N-1} \tau_k \sum_{j<k} \tau_j = \frac{1}{2}\left[\left(\sum_{k=0}^{N-1} \tau_k\right)^2 - \sum_{k=0}^{N-1} \tau_k^2\right],$$

then combining the two terms involving $\sum_{k=0}^{N-1} \tau_k^2$ yields (21). □

G Proof of Corollary 3.2

Proof. By Theorem 3.1, (21) holds. We show that all terms involving the inexactness sequence vanish as $N \to \infty$. First, since $\sum_{k=0}^{\infty} \tau_k^2 < \infty$, for every $\varepsilon > 0$ there exists $K \in \mathbb{N}_0$ such that

$$\sum_{k=K}^{\infty} \tau_k^2 < \varepsilon^2.$$

Splitting the weighted sum at K and applying the Cauchy–Schwarz inequality to its tail gives

$$\begin{aligned}\frac{1}{N}\sum_{k=0}^{N-1} \tau_k \sqrt{W_0^2 + 2L_\gamma n\eta^2 k} \le \frac{1}{N}\sum_{k=0}^{K-1} \tau_k \sqrt{W_0^2 + 2L_\gamma n\,\eta^2 k} \\ + \left(\sum_{k=K}^{\infty} \tau_k^2\right)^{1/2} \left[\frac{1}{N^2}\sum_{k=K}^{N-1} \left(W_0^2 + 2L_\gamma n\,\eta^2 k\right)\right]^{1/2}.\end{aligned}$$

The first term tends to zero for fixed K, while

$$\frac{1}{N^2}\sum_{k=K}^{N-1} \left(W_0^2 + 2L_\gamma n\,\eta^2 k\right),$$

remains bounded as $N \to \infty$. Hence, taking first $N \to \infty$ and then $\varepsilon \downarrow 0$ yields

$$\frac{1}{N}\sum_{k=0}^{N-1} \tau_k \sqrt{W_0^2 + 2L_\gamma n\,\eta^2 k} \to 0.$$

Next, square summability also implies

$$\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1} \tau_k \to 0.$$

Indeed, for the same fixed K,

$$\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}\tau_k \le \frac{1}{\sqrt{N}}\sum_{k=0}^{K-1}\tau_k + \left(\sum_{k=K}^{\infty}\tau_k^2\right)^{1/2},$$

i.e.,

$$\frac{1}{N}\left(\sum_{k=0}^{N-1}\tau_k\right)^2 \to 0,$$

leading to

$$\frac{1}{N}\sum_{k=0}^{N-1}\tau_k^2 \le \frac{1}{N}\sum_{k=0}^{\infty}\tau_k^2 \to 0.$$

Since also $W_0^2/(2\eta N) \to 0$, taking the upper limit in the bound of Theorem 3.1 gives

$$\limsup_{N\to\infty} D_{\mathrm{KL}}(\nu_N\|\pi_\gamma) \le L_\gamma n\,\eta.$$

Applying Pinsker's inequality ensures

$$\mathrm{TV}(\nu_N,\pi_\gamma) \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(\nu_N\|\pi_\gamma)}.$$

Combining this with the triangle inequality for total variation yields

$$\mathrm{TV}(\nu_N,\pi) \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(\nu_N\|\pi_\gamma)} + \mathrm{TV}(\pi_\gamma,\pi),$$

giving (23). □

H Proof of Theorem 3.3

Let $\mathcal{F}_k$ denote the information available before the kth inner solution, and let ξ_k collect the randomness used by the inner solver. We define

$$\mathcal{G}_k := \mathcal{F}_k \vee \sigma(\xi_k),$$

where $\sigma(\xi_k)$ denotes the σ-algebra generated by ξ_k and $\vee$ denotes the smallest σ-algebra containing both $\mathcal{F}_k$ and $\sigma(\xi_k)$.

Proof. From L_γ-smoothness of $\mathcal{U}_\gamma$ and (20), for every $k \in \mathbb{N}_0$,

$$\begin{aligned}\mathcal{U}_\gamma(x_{k+1}) &\le \mathcal{U}_\gamma(x_k) + \langle\nabla\mathcal{U}_\gamma(x_k), x_{k+1}-x_k\rangle + \frac{L_\gamma}{2}\|x_{k+1}-x_k\|^2 \\ &= \mathcal{U}_\gamma(x_k) + \left\langle\nabla\mathcal{U}_\gamma(x_k), -\eta\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k) + \sqrt{2\eta}\,\mathcal{Z}_k\right\rangle + \frac{L_\gamma}{2}\left\|-\eta\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k) + \sqrt{2\eta}\,\mathcal{Z}_k\right\|^2\end{aligned}$$

We now take conditional expectation with respect to $\mathcal{G}_k$. Under this conditioning, x_k and the inexact oracle are fixed, while $\mathcal{Z}_k$ is centered and independent by construction, i.e.,

$$\begin{aligned}\mathbb{E}[\mathcal{U}_\gamma(x_{k+1}) \mid \mathcal{G}_k] \le \mathcal{U}_\gamma(x_k) &+ \mathbb{E}\left[\left\langle\nabla\mathcal{U}_\gamma(x_k), -\eta\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k) + \sqrt{2\eta}\,\mathcal{Z}_k\right\rangle\middle|\mathcal{G}_k\right] \\ &+ \frac{L_\gamma}{2}\mathbb{E}\left[\left\|-\eta\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k) + \sqrt{2\eta}\,\mathcal{Z}_k\right\|^2\middle|\mathcal{G}_k\right].\end{aligned}$$

For the linear term, it holds that

$$\begin{aligned}\mathbb{E}\left[\left\langle\nabla\mathcal{U}_\gamma(x_k), -\eta\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k) + \sqrt{2\eta}\,\mathcal{Z}_k\right\rangle\middle|\mathcal{G}_k\right] &= -\eta\left\langle\nabla\mathcal{U}_\gamma(x_k), \nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\rangle \\ &\quad + \sqrt{2\eta}\,\left\langle\nabla\mathcal{U}_\gamma(x_k), \mathbb{E}[\mathcal{Z}_k \mid \mathcal{G}_k]\right\rangle.\end{aligned}$$

Since $\mathbb{E}[\mathcal{Z}_k \mid \mathcal{G}_k] = 0$, this reduces to

$$\mathbb{E}\left[\left\langle\nabla\mathcal{U}_\gamma(x_k), -\eta\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k) + \sqrt{2\eta}\,\mathcal{Z}_k\right\rangle\middle|\mathcal{G}_k\right] = -\eta\left\langle\nabla\mathcal{U}_\gamma(x_k), \nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\rangle.$$

For the quadratic term, expanding the square and taking conditional expectation gives

$$\mathbb{E}\left[\left\|-\eta\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)+\sqrt{2\eta}\,\mathcal{Z}_k\right\|^2\middle|\mathcal{G}_k\right]=\eta^2\left\|\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\|^2+2\eta\,\mathbb{E}[\|\mathcal{Z}_k\|^2\mid\mathcal{G}_k]\\-2\eta\sqrt{2\eta}\,\left\langle\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k),\mathbb{E}[\mathcal{Z}_k\mid\mathcal{G}_k]\right\rangle.$$

Again, $\mathbb{E}[\mathcal{Z}_k\mid\mathcal{G}_k]=0$, i.e., the mixed term vanishes. Moreover, since $\mathcal{Z}_k\sim\mathcal{N}(0,I_n)$,

$$\mathbb{E}[\|\mathcal{Z}_k\|^2\mid\mathcal{G}_k]=\mathbb{E}[\|\mathcal{Z}_k\|^2]=n,$$

i.e.,

$$\mathbb{E}\left[\left\|-\eta\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)+\sqrt{2\eta}\,\mathcal{Z}_k\right\|^2\middle|\mathcal{G}_k\right]=\eta^2\left\|\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\|^2+2\eta n.$$

Substituting the linear and quadratic expectations into the previous inequality yields

$$\mathbb{E}[\mathcal{U}_\gamma(x_{k+1})\mid\mathcal{G}_k]\le\mathcal{U}_\gamma(x_k)-\eta\left\langle\nabla\mathcal{U}_\gamma(x_k),\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\rangle+\frac{L_\gamma}{2}\big(\eta^2\left\|\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\|^2+2\eta n\big),$$

that is,

$$\mathbb{E}[\mathcal{U}_\gamma(x_{k+1})\mid\mathcal{G}_k]\le\mathcal{U}_\gamma(x_k)-\eta\left\langle\nabla\mathcal{U}_\gamma(x_k),\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\rangle+\frac{L_\gamma\eta^2}{2}\left\|\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\|^2+\eta L_\gamma n.$$

Now, from (10), it holds that $\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)=\nabla\mathcal{U}_\gamma(x_k)+e_k$, i.e.,

$$-\left\langle\nabla\mathcal{U}_\gamma(x_k),\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\rangle=-\left\|\nabla\mathcal{U}_\gamma(x_k)\right\|^2-\left\langle\nabla\mathcal{U}_\gamma(x_k),e_k\right\rangle.$$

Applying the Cauchy–Schwarz and Young's inequalities ensures

$$-\left\langle\nabla\mathcal{U}_\gamma(x_k),e_k\right\rangle\le\left\|\nabla\mathcal{U}_\gamma(x_k)\right\|\,\|e_k\|\le\frac{1}{2}\left\|\nabla\mathcal{U}_\gamma(x_k)\right\|^2+\frac{1}{2}\|e_k\|^2,$$

i.e.,

$$-\left\langle\nabla\mathcal{U}_\gamma(x_k),\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\rangle\le-\frac{1}{2}\left\|\nabla\mathcal{U}_\gamma(x_k)\right\|^2+\frac{1}{2}\|e_k\|^2.$$

In addition, we have

$$\left\|\nabla\mathcal{U}_\gamma^{\varepsilon_k}(x_k)\right\|^2=\left\|\nabla\mathcal{U}_\gamma(x_k)+e_k\right\|^2\le2\left\|\nabla\mathcal{U}_\gamma(x_k)\right\|^2+2\|e_k\|^2.$$

This consequently implies that

$$\mathbb{E}[\mathcal{U}_\gamma(x_{k+1})\mid\mathcal{G}_k]\le\mathcal{U}_\gamma(x_k)-\eta\left(\frac{1}{2}-L_\gamma\eta\right)\left\|\nabla\mathcal{U}_\gamma(x_k)\right\|^2+\eta\left(\frac{1}{2}+L_\gamma\eta\right)\|e_k\|^2+\eta L_\gamma n.$$

Taking conditional expectation with respect to $\mathcal{F}_k$, using $\mathbb{E}[\delta_k^2\mid\mathcal{F}_k]\le\delta^2$, and then taking total expectation ensure

$$\mathbb{E}[\mathcal{U}_\gamma(x_{k+1})]\le\mathbb{E}[\mathcal{U}_\gamma(x_k)]-\eta\left(\frac{1}{2}-L_\gamma\eta\right)\mathbb{E}\left\|\nabla\mathcal{U}_\gamma(x_k)\right\|^2+\eta\left(\frac{1}{2}+L_\gamma\eta\right)\delta^2+\eta L_\gamma n.\quad(49)$$

By the Polyak–Łojasiewicz inequality (24),

$$\left\|\nabla\mathcal{U}_\gamma(x)\right\|^2\ge2\mu_\gamma(\mathcal{U}_\gamma(x)-\mathcal{U}_\gamma^\star),$$

for $\mathcal{U}_\gamma^\star:=\inf\mathcal{U}_\gamma$, i.e.,

$$\eta\left(\frac{1}{2}-L_\gamma\eta\right)\left\|\nabla\mathcal{U}_\gamma(x_k)\right\|^2\ge2\mu_\gamma\eta\left(\frac{1}{2}-L_\gamma\eta\right)\big(\mathcal{U}_\gamma(x_k)-\mathcal{U}_\gamma^\star\big)=\mu_\gamma\eta(1-2L_\gamma\eta)\big(\mathcal{U}_\gamma(x_k)-\mathcal{U}_\gamma^\star\big).$$

Substituting this lower bound into (49) and subtracting $\mathcal{U}_\gamma^\star$ from both sides gives

$$\mathbb{E}[\mathcal{U}_\gamma(x_{k+1})-\mathcal{U}_\gamma^\star]\le(1-\mu_\gamma\eta(1-2L_\gamma\eta))\,\mathbb{E}[\mathcal{U}_\gamma(x_k)-\mathcal{U}_\gamma^\star]+\eta\left(\frac{1}{2}+L_\gamma\eta\right)\delta^2+\eta L_\gamma n,$$

which leads to the inequality (26).

Now, set $a_k := \mathbb{E}[\mathcal{U}_\gamma(x_k) - \mathcal{U}_\gamma^\star]$ and $b := \eta\left(\frac{1}{2} + L_\gamma \eta\right)\delta^2 + \eta L_\gamma n$. From (26), we obtain

$$a_{k+1} \le \rho_{\mathrm{PL}} a_k + b,$$

for ρ_{PL} defined in (25). Iterating this recursion yields

$$a_k \le \rho_{\mathrm{PL}}^k a_0 + b \sum_{j=0}^{k-1} \rho_{\mathrm{PL}}^j = \rho_{\mathrm{PL}}^k a_0 + b \frac{1 - \rho_{\mathrm{PL}}^k}{1 - \rho_{\mathrm{PL}}}.$$

Since $1 - \rho_{\mathrm{PL}}^k \le 1$, it is clear that

$$a_k \le \rho_{\mathrm{PL}}^k a_0 + \frac{b}{1 - \rho_{\mathrm{PL}}}.$$

Substituting back the definitions of a_k, a_0, and b results in

$$\mathbb{E}[\mathcal{U}_\gamma(x_k) - \mathcal{U}_\gamma^\star] \le \rho_{\mathrm{PL}}^k \mathbb{E}\left[\mathcal{U}_\gamma(x_0) - \mathcal{U}_\gamma^\star\right] + \frac{\eta\left(\frac{1}{2} + L_\gamma \eta\right)\delta^2 + \eta L_\gamma n}{1 - \rho_{\mathrm{PL}}},$$

adjusting the inequalities (27) and (28). □

I Proof of Theorem 3.4

Proof. The proof follows the same smoothness calculation as Theorem 3.3, with conditioning first on $\mathcal{G}_k$ and then on $\mathcal{F}_k$. The adaptive bound $\mathbb{E}[\delta_k^2 \mid \mathcal{F}_k] \le \tau_k^2$ implies

$$\mathbb{E}[\|e_k\|^2 \mid \mathcal{F}_k] \le \tau_k^2,$$

instead of the uniform bound by δ^2. This consequently implies

$$\mathbb{E}[\mathcal{U}_\gamma(x_{k+1}) - \mathcal{U}_\gamma^\star] \le \rho_{\mathrm{PL}}\, \mathbb{E}[\mathcal{U}_\gamma(x_k) - \mathcal{U}_\gamma^\star] + \eta\left(\frac{1}{2} + L_\gamma \eta\right)\tau_k^2 + \eta L_\gamma n.$$

Iterating the recursion leads to the inequality (28). □

J Inexact iPULA chain vs. exact Moreau–ULA chain

The following result compares the inexact iPULA chain with the exact Moreau–ULA chain under a synchronous coupling, i.e., when both chains are driven by the same Gaussian noise. This bound isolates the additional sampling error induced by the inexact proximal computation.

Theorem J.1 (Fixed-tolerance Wasserstein transfer). *Suppose Assumption 2.1 holds, $\mathcal{U}$ is σ-strongly convex, and there exists $\delta \ge 0$ such that $\delta_k \le \delta$ almost surely for all $k \in \mathbb{N}_0$. Let $\{\overline{x}_k\}_{k\in\mathbb{N}_0}$ and $\{x_k\}_{k\in\mathbb{N}_0}$ be the exact and inexact chains (19) and (20), respectively, driven by the same Gaussian noises and started from the same initial point. Then, for every $k \in \mathbb{N}_0$,*

$$\mathcal{W}_2\left(\mathcal{L}(x_k), \mathcal{L}(\overline{x}_k)\right) \le \frac{\eta\delta(1 - q^k)}{1 - q}. \tag{50}$$

Consequently,

$$\mathcal{W}_2\left(\mathcal{L}(x_k), \pi_\gamma\right) \le \mathcal{W}_2\left(\mathcal{L}(\overline{x}_k), \pi_\gamma\right) + \frac{\eta\delta(1 - q^k)}{1 - q}. \tag{51}$$

Proof. Let us couple the exact and inexact chains using the same Gaussian noises, i.e.,

$$\begin{aligned}
\overline{x}_{k+1} &= \overline{x}_k - \eta \nabla \mathcal{U}_\gamma(\overline{x}_k) + \sqrt{2\eta}\, \mathcal{Z}_k, \\
x_{k+1} &= x_k - \eta\left(\nabla \mathcal{U}_\gamma(x_k) + e_k\right) + \sqrt{2\eta}\, \mathcal{Z}_k.
\end{aligned}$$

Set $d_k := x_k - \overline{x}_k$, i.e.,

$$d_{k+1} = d_k - \eta(\nabla \mathcal{U}_\gamma(x_k) - \nabla \mathcal{U}_\gamma(\overline{x}_k)) - \eta e_k.$$

For convenience, let us define

$$A_k := \nabla \mathcal{U}_\gamma(x_k) - \nabla \mathcal{U}_\gamma(\overline{x}_k),$$

i.e., $d_{k+1} = d_k - \eta A_k - \eta e_k$. We first estimate the deterministic part $d_k - \eta A_k$. Expanding its squared norm gives

$$\|d_k - \eta A_k\|^2 = \|d_k\|^2 - 2\eta \langle d_k, A_k \rangle + \eta^2 \|A_k\|^2 .$$

Since $\mathcal{U}_\gamma$ is m_γ-strongly convex, its gradient is m_γ-strongly monotone, leading to

$$\langle d_k, A_k \rangle = \langle x_k - \overline{x}_k, \nabla \mathcal{U}_\gamma(x_k) - \nabla \mathcal{U}_\gamma(\overline{x}_k) \rangle \geq m_\gamma \|d_k\|^2 .$$

Moreover, since $\mathcal{U}_\gamma$ is L_γ-smooth,

$$\|A_k\| = \|\nabla \mathcal{U}_\gamma(x_k) - \nabla \mathcal{U}_\gamma(\overline{x}_k)\| \leq L_\gamma \|d_k\| .$$

Substituting these two bounds into the expansion above yields

$$\|d_k - \eta A_k\|^2 \leq \|d_k\|^2 - 2m_\gamma \eta \|d_k\|^2 + L_\gamma^2 \eta^2 \|d_k\|^2 = (1 - 2m_\gamma \eta + L_\gamma^2 \eta^2) \|d_k\|^2 .$$

Define $q := \sqrt{1 - 2m_\gamma \eta + L_\gamma^2 \eta^2}$. Under the step-size condition of the theorem, we have $q \in [0, 1)$, and consequently it implies

$$\|d_k - \eta A_k\| \leq q \|d_k\| .$$

It follows from the triangle inequality that

$$\|d_{k+1}\| \leq \|d_k - \eta A_k\| + \eta \|e_k\| \leq q \|d_k\| + \eta \|e_k\| ,$$

almost surely. By the assumption that $\delta_k \leq \delta$ almost surely for all $k \in \mathbb{N}_0$, we have $\|e_k\| \leq \delta$ almost surely for all k, which leads to

$$\|d_{k+1}\| \leq q \|d_k\| + \eta\delta.$$

Iterating this inequality, we come to

$$\|d_k\| \leq q^k \|d_0\| + \eta\delta \sum_{j=0}^{k-1} q^j .$$

Since the two chains start from the same point $d_0 = 0$, it can be deduced that

$$\|d_k\| \leq \eta\delta \sum_{j=0}^{k-1} q^j = \frac{\eta\delta(1 - q^k)}{1 - q},$$

almost surely. Using the fact that $(x_k, \overline{x}_k)$ is a coupling of laws $\mathcal{L}(x_k)$ and $\mathcal{L}(\overline{x}_k)$, the coupling characterization of $\mathcal{W}_2$ yields

$$\mathcal{W}_2 \left(\mathcal{L}(x_k), \mathcal{L}(\overline{x}_k)\right) \leq \left(\mathbb{E} \|x_k - \overline{x}_k\|^2\right)^{1/2} = \left(\mathbb{E} \|d_k\|^2\right)^{1/2} \leq \frac{\eta\delta(1 - q^k)}{1 - q},$$

which leads to (50). Finally, applying the triangle inequality for $\mathcal{W}_2$ ensures

$$\mathcal{W}_2 \left(\mathcal{L}(x_k), \pi_\gamma\right) \leq \mathcal{W}_2 \left(\mathcal{L}(x_k), \mathcal{L}(\overline{x}_k)\right) + \mathcal{W}_2 \left(\mathcal{L}(\overline{x}_k), \pi_\gamma\right),$$

and the substitution of this into the previous estimate leads to the inequality (51). □

The following result provides the analog of Theorem J.1 under a time-varying proximal accuracy schedule. Rather than accumulating a fixed inexactness level, the resulting transfer error takes the form of a geometrically weighted convolution of the tolerances τ_k.

Theorem J.2 (Adaptive-tolerance Wasserstein transfer). *Suppose Assumption 2.1 holds, $\mathcal{U}$ is σ-strongly convex, and there exists a deterministic sequence $\{\tau_k\}_{k \geq 0}$ such that $\delta_k \leq \tau_k$ almost surely for all $k \in \mathbb{N}_0$. Then, under the same synchronous coupling as in Theorem J.1, for every $k \in \mathbb{N}_0$,*

$$\mathcal{W}_2 \left(\mathcal{L}(x_k), \mathcal{L}(\overline{x}_k)\right) \leq \eta \sum_{j=0}^{k-1} q^{k-1-j} \tau_j, \tag{52}$$

where q is given by (29). Consequently,

$$\mathcal{W}_2 \left(\mathcal{L}(x_k), \pi_\gamma\right) \leq \mathcal{W}_2 \left(\mathcal{L}(\overline{x}_k), \pi_\gamma\right) + \eta \sum_{j=0}^{k-1} q^{k-1-j} \tau_j. \tag{53}$$

Proof. The proof follows the same synchronous coupling argument as in Appendix J. Starting from $\|d_{k+1}\| \le q\,\|d_k\| + \eta\delta_k$ and obtaining $\|d_{k+1}\| \le q\,\|d_k\| + \eta\tau_k$. Iterating this recursion yields

$$\|d_k\| \le \eta \sum_{j=0}^{k-1} q^{k-1-j}\tau_j,$$

since $d_0 = 0$. The claimed Wasserstein bounds follow exactly as before. □

The previous Wasserstein transfer bounds compare the inexact chain with the corresponding exact Moreau-smoothed chain.

K Proof of Theorem 3.5

Proof. By the triangle inequality, Theorem J.2, and the exact-LMC estimate (30),

$$\begin{aligned}\mathcal{W}_2(\mathcal{L}(x_k), \pi_\gamma) &\le \mathcal{W}_2(\mathcal{L}(x_k), \mathcal{L}(\overline{x}_k)) + \mathcal{W}_2(\mathcal{L}(\overline{x}_k), \pi_\gamma) \\ &\le \eta \sum_{j=0}^{k-1} q^{k-1-j}\tau_j + (1 - m_\gamma\eta)^k \mathcal{W}_2(\mathcal{L}(x_0), \pi_\gamma) + 1.65\frac{L_\gamma}{m_\gamma}\sqrt{\eta n},\end{aligned}$$

which is (31). The fixed-tolerance form (32) follows by setting $\tau_j \equiv \delta$ and summing the geometric series. □

L Consequences of adaptive tolerance schedules

We collect here two direct consequences of the adaptive objective-gap and Wasserstein transfer bounds. These results clarify how common choices of lower-level proximal tolerances affect the upper-level error estimates.

Corollary L.1 (Step-size-matched tolerance rule). *Suppose the assumptions of Theorem 3.4 hold. Assume that the tolerance schedule satisfies*

$$\tau_k \le c\sqrt{\eta}, \quad \forall k \in \mathbb{N}_0,$$

for some constant $c > 0$. Then, for every $k \in \mathbb{N}_0$,

$$\mathbb{E}\left[\mathcal{U}_\gamma(x_k) - \mathcal{U}_\gamma(x^\star)\right] \le \rho_{\mathrm{PL}}^k \mathbb{E}\left[\mathcal{U}_\gamma(x_0) - \mathcal{U}_\gamma(x^\star)\right] + \frac{\eta L_\gamma n + c^2\eta^2(\frac{1}{2} + L_\gamma\eta)}{1 - \rho_{\mathrm{PL}}}. \tag{54}$$

Equivalently, we have

$$\mathbb{E}\left[\mathcal{U}_\gamma(x_k) - \mathcal{U}_\gamma(x^\star)\right] \le \rho_{\mathrm{PL}}^k \mathbb{E}\left[\mathcal{U}_\gamma(x_0) - \mathcal{U}_\gamma(x^\star)\right] + \frac{L_\gamma n + c^2\eta(\frac{1}{2} + L_\gamma\eta)}{\mu_\gamma(1 - 2L_\gamma\eta)}. \tag{55}$$

The expanded form (55) separates the two sources of error. Since $1 - \rho_{\mathrm{PL}} = \mu_\gamma\eta(1 - 2L_\gamma\eta)$, the Gaussian-noise contribution satisfies

$$\frac{\eta L_\gamma n}{1 - \rho_{\mathrm{PL}}} = \frac{L_\gamma n}{\mu_\gamma(1 - 2L_\gamma\eta)} = O(1),$$

whereas the step-size-matched oracle contribution satisfies

$$\frac{c^2\eta^2(\frac{1}{2} + L_\gamma\eta)}{1 - \rho_{\mathrm{PL}}} = \frac{c^2\eta(\frac{1}{2} + L_\gamma\eta)}{\mu_\gamma(1 - 2L_\gamma\eta)} = O(\eta).$$

Thus, the rule $\tau_k = O(\sqrt{\eta})$ makes the inexact-proximal contribution vanish as $\eta \to 0$, while the Gaussian-noise contribution remains nonzero in this objective-gap-to-minimizer estimate. This is expected because iPULA is a sampling method for π_γ, not a deterministic optimization method converging to $x^\star$.

Corollary L.2 (Adaptive schedules). *Under the assumptions of Theorem J.2:*

(a) *If $\tau_k \equiv \tau$ for all $k \in \mathbb{N}_0$, then*

$$\mathcal{W}_2\left(\mathcal{L}(x_k), \mathcal{L}(\overline{x}_k)\right) \leq \frac{\eta\tau(1-q^k)}{1-q},$$

which recovers the fixed-tolerance transfer bound with $\delta = \tau$;

(b) *if $\tau_k \to 0$, then the inexactness perturbation term in (53) vanishes asymptotically;*

(c) *if $\tau_k \leq c\sqrt{\eta}$ for all $k \in \mathbb{N}_0$, then*

$$\mathcal{W}_2\left(\mathcal{L}(x_k), \mathcal{L}(\overline{x}_k)\right) \leq \frac{c\,\eta^{3/2}(1-q^k)}{1-q}.$$

Corollary L.2 clarifies how different lower-level tolerance rules affect the Wasserstein transfer error. A fixed tolerance produces a persistent inexactness bias, a vanishing tolerance removes the additional inexactness perturbation asymptotically, and the step-size-matched rule $\tau_k = O(\sqrt{\eta})$ yields an inexactness contribution of order $\eta^{3/2}$ before normalization by $1-q$.