Stochastic Dynamic Barrier Perturbed Gradient Methods for Nonconvex Simple Bilevel Optimization

Mohammad Mahdi Ahmadi[1], **Jincheng Cao**[2], **Aryan Mokhtari**[2], and **Erfan Yazdandoost Hamedani**[3]

[1]Industrial and Systems Engineering, Mississippi State University
[2]Electrical and Computer Engineering, The University of Texas at Austin
[3]Systems and Industrial Engineering, The University of Arizona







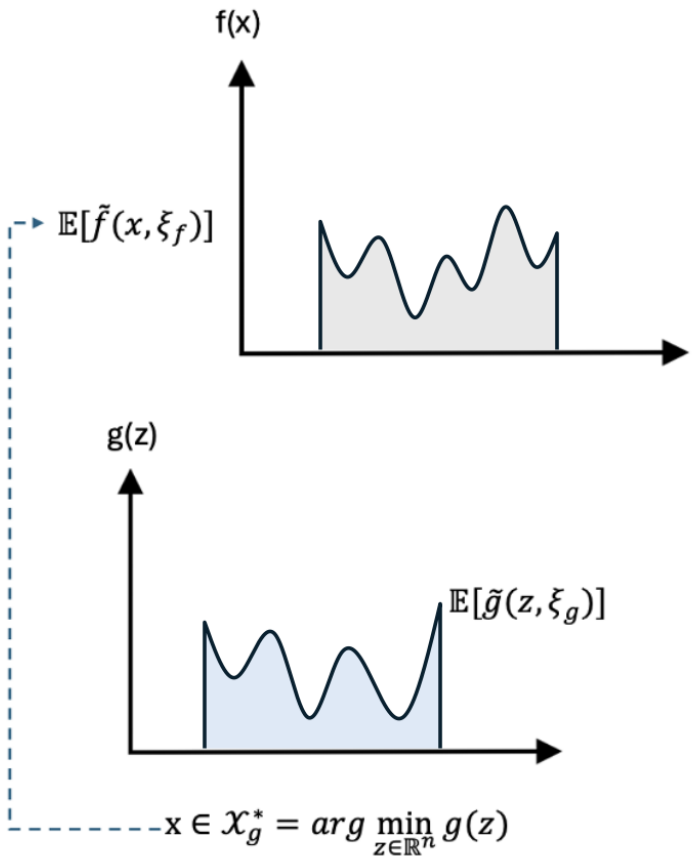


Problem Statement

cont. diff.
nonconvex

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) = \mathbb{E}\big[\tilde{f}(\mathbf{x},\xi_f)\big]$$

$$\text{s.t.} \quad \mathbf{x} \in \mathcal{X}_g^* := \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^n} \; g(\mathbf{z}) = \mathbb{E}\big[\tilde{g}(\mathbf{z},\xi_g)\big]$$



- Minimize f over the solution set of another problem g
- Both f and g are nonconvex and only accessible through stochastic gradients
- **Application: LLM unlearning**
  - lower level g: forget loss (remove target data)
  - upper level f: retain loss (preserve utility)
- **Goal:** single-loop stochastic method with joint upper- and lower-level stationarity guarantees

Related Work

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) = \mathbb{E}[\tilde{f}(\mathbf{x}, \xi_f)] \quad \text{s.t.} \quad \mathbf{x} \in \arg\min_{\mathbf{z}\in\mathbb{R}^n} \; g(\mathbf{z}) = \mathbb{E}[\tilde{g}(\mathbf{z}, \xi_g)]. \qquad \text{(SBO)}$$

1) **Convex–convex** [Jalilzadeh et al. (2024)]
2) **Nonconvex–convex** [Cao et al. (2023), Giang-Tran et al. (2025)] g convex
3) **Deterministic nonconvex–nonconvex** [Reisizadeh et al. (2026), Cao et al. (2026)] exact gradients
4) **Stochastic nonconvex–nonconvex** [Hsieh et al. (2024)] lower level only

Gap: No joint (ϵ_f, ϵ_g)-stationarity with explicit sample complexity in the stochastic nonconvex–nonconvex setting.

Why is this hard?

- Feasible set $\mathcal{X}_g^*$ is implicit: projection, Tikhonov regularization, and value-function reformulations all break down for nonconvex, stochastic g

- **Dynamic Barrier Gradient Descent (DBGD)** [Gong et al.'21]: combine $\nabla f, \nabla g$ via a dual multiplier λ_k (no projection)

$$\lambda_k = \max\left\{\beta_k - \frac{\nabla g_k^\top \nabla f_k}{\|\nabla g_k\|^2}, 0\right\}$$

$$\mathbf{d}_k = \nabla f_k + \lambda_k \nabla g_k, \qquad \mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{d}_k$$

- Deterministic rate [Cao et al.'26]: $\mathcal{O}(\max\{\epsilon_f^{-2}, \epsilon_g^{-2}\})$ — exact gradients only

Key challenge: $\lambda_k \propto \|\nabla g(\mathbf{x}_k)\|^{-1}$ diverges near lower-level stationarity.

Under noise, $\tilde{\lambda}_k$ (ratio of correlated mini-batch gradients) is biased and lacks bounded variance.

Proposed Method: SDBPG

Stochastic DBGD with dual perturbation

- **Idea:** perturb the denominator with $\gamma_k \geq 0$ (perturbed dual problem)

$$\lambda_{\gamma,k} = \max\left\{\beta_k - \frac{\nabla g_k^\top \nabla f_k}{\|\nabla g_k\|^2 + \gamma_k}, 0\right\}$$

- **Stochastic estimator** (single-loop, mini-batch gradients)

$$\tilde{\lambda}_{\gamma,k} = \max\left\{\beta_k - \frac{\nabla \tilde{g}_k^\top \nabla \tilde{f}_k}{\|\nabla \tilde{g}_k\|^2 + \gamma_k}, 0\right\}$$

$$\tilde{\mathbf{d}}_k = \nabla \tilde{f}_k + \tilde{\lambda}_{\gamma,k} \nabla \tilde{g}_k, \qquad \mathbf{x}_{k+1} = \mathbf{x}_k - \eta\, \tilde{\mathbf{d}}_k$$

- $\tilde{\lambda}_{\gamma,k}$ is biased $\Rightarrow$ we analyze the direction $\tilde{\mathbf{d}}_k$ instead, via a joint Lipschitz analysis of $T_\gamma(u,v) = \frac{\langle u,v\rangle}{\|v\|^2+\gamma} v$
- **But:** matching the deterministic rate needs γ_k and batch sizes tied to the true $\|\nabla g(\mathbf{x}_k)\|$ $\Rightarrow$ not implementable

The Rare-Visit Condition

- $\lambda_{\gamma,k}$ is large only when ∇g is small *and* negatively aligned with ∇f — the **bad region**:

$$\mathcal{B}(\tau,\theta) \triangleq \left\{\mathbf{x}:\ \|\nabla g(\mathbf{x})\|^2 \le \tau,\quad \nabla g(\mathbf{x})^\top \nabla f(\mathbf{x}) < -\theta\|\nabla g(\mathbf{x})\|^2\right\}$$

- **Rare-visit assumption:** expected number of visits is sublinear in K

$$\textstyle\sum_{k=0}^{K-1} \mathbb{P}\big(\mathbf{x}_k \in \mathcal{B}(\tau,\theta)\big) = \mathcal{O}(K^{1-\varsigma}), \qquad \varsigma \in (0,1]$$

- **Bad-region detector:** with independent batches $\nabla \hat{f}_k, \nabla \hat{g}_k$, if

$$\|\nabla \hat{g}_k\|^2 \le \tau/2 \ \text{ and } \ \nabla \hat{g}_k^\top \nabla \hat{f}_k < -\theta\|\nabla \hat{g}_k\|^2 \quad \Rightarrow \quad \tilde{\lambda}_{\gamma,k} \leftarrow 0$$

  - follow the upper-level gradient $\Rightarrow$ by negative alignment, pushes iterates out of $\mathcal{B}(\tau,\theta)$
  - fixed γ, N_f, N_g — no dependence on $\|\nabla g(\mathbf{x}_k)\|$

Convergence of SDBPG

Definition [Cao et al.'26]

$\hat{\mathbf{x}}$ is an (ϵ_f, ϵ_g)-stationary point if there exists $\lambda \geq 0$ such that

$$\|\nabla g(\hat{\mathbf{x}})\|^2 \leq \epsilon_g \qquad \text{and} \qquad \|\nabla f(\hat{\mathbf{x}}) + \lambda \nabla g(\hat{\mathbf{x}})\|^2 \leq \epsilon_f.$$

Theorem 3.3 (SDBPG)

Under the rare-visit assumption, with $\eta = \beta = \theta = K^{-1/4}$, $\tau = K^{-1/2}$, $\gamma = K^{-1}$, $N_f = K$, $N_g = K^2$, there exists k^* such that

$$\mathbb{E}\|\mathbf{d}_{k^*}\|^2, \quad \mathbb{E}\|\nabla g(\mathbf{x}_{k^*})\|^2 \leq \mathcal{O}\big(K^{-1/2} + K^{1/2-\varsigma}\big).$$

For $\varsigma \in (\frac{1}{2}, 1]$, $\delta \triangleq \varsigma - \frac{1}{2}$: upper-level $\mathcal{O}(\epsilon^{-2/\delta})$, lower-level $\mathcal{O}(\epsilon^{-3/\delta})$ samples.

If $\mathcal{B}(\tau, \theta) = \emptyset$: $\mathcal{O}(\epsilon^{-4})$ and $\mathcal{O}(\epsilon^{-6})$, recovering the deterministic rate $\mathcal{O}(K^{-1/2})$.

Removing the Rare-Visit Assumption

- **PR-SDBPG:** replace the constrained subproblem with a penalty-regularized one $\Rightarrow$ closed form $\mathbf{d}_k = \nabla f_k + \hat{\lambda}_k \nabla g_k$,

$$\hat{\lambda}_k = \left[\frac{\mu \left(\beta(\|\nabla g_k\|^2 + \gamma) - \nabla g_k^\top \nabla f_k \right)}{(1+\mu)\|\nabla g_k\|^2 + \gamma \;\geq\; \gamma > 0} \right]_+$$

  - Uniformly bounded multiplier; fixed γ and batch sizes

- **VR-PR-SDBPG:** add STORM-type momentum trackers v_f, v_g [Cutkosky&Orabona'19]

Theorems 4.2 & 4.3 Without the rare-visit assumption, after $K = \mathcal{O}(\epsilon^{-4})$ iterations, $\exists\, k^*$ with $\mathbb{E}\|\mathbf{d}_{k^*}\|^2,\ \mathbb{E}\|\nabla g(\mathbf{x}_{k^*})\|^2 \leq \mathcal{O}(\epsilon)$:

PR-SDBPG: upper $\mathcal{O}(\epsilon^{-8})$, lower $\mathcal{O}(\epsilon^{-10})$

VR-PR-SDBPG (+ mean-square smoothness)**:** upper $\mathcal{O}(\epsilon^{-6})$, lower $\mathcal{O}(\epsilon^{-8})$

Comparison with Other Studies

Reference	Type	f	g	Upper	Lower	Extra
Jalilzadeh et al. (2024)	S	C	C	$\mathcal{O}(\epsilon^{-4})$		convexity
Cao et al. (2023)	S	NC	C	$\tilde{\mathcal{O}}(\epsilon^{-3})$		g convex
Giang-Tran et al. (2025)	S	NC	C	$\tilde{\mathcal{O}}(\epsilon^{-7})$		g convex
Hsieh et al. (2024)	S	NC	NC	–	$\mathcal{O}(\epsilon_g^{-4})$	lower only
Reisizadeh et al. (2026)	D	NC	NC	N/A (deterministic)		–
Cao et al. (2026)	D	NC	NC	N/A (deterministic)		–
SDBPG*	S	NC	NC	$\mathcal{O}(\epsilon^{-2/\delta})$	$\mathcal{O}(\epsilon^{-3/\delta})$	rare-visit, $\delta \in (0, \frac{1}{2}]$
PR-SDBPG*	S	NC	NC	$\mathcal{O}(\epsilon^{-8})$	$\mathcal{O}(\epsilon^{-10})$	**none**
VR-PR-SDBPG*	S	NC	NC	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{O}(\epsilon^{-8})$	**none**†

S/D: stochastic/deterministic **C/NC**: convex/nonconvex

SDBPG with $\delta = \frac{1}{2}$ (e.g., empty bad region): $\mathcal{O}(\epsilon^{-4})$, $\mathcal{O}(\epsilon^{-6})$

†no rare-visit assumption; uses standard mean-square smoothness for STORM

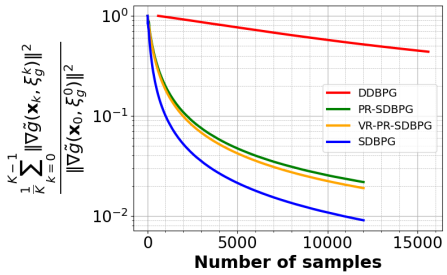
Numerical Experiments: LLM Unlearning

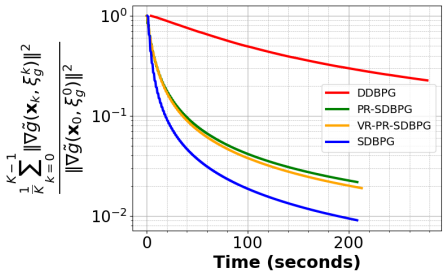
Setup: f = retain loss (CE), g = forget loss (NPO); TOFU (Llama-3.2-1B), MUSE-News (Llama-2-7B)

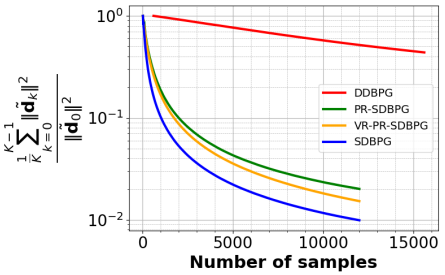
SDBPG, **PR-SDBPG**, **VR-PR-SDBPG** vs. deterministic **DDBPG**

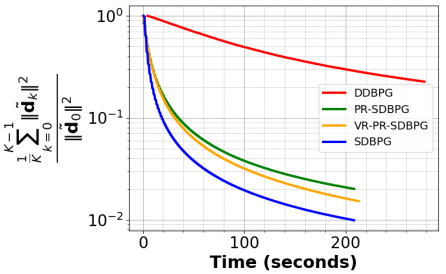
$\|\nabla \tilde{g}\|^2$ vs. samples | $\|\nabla \tilde{g}\|^2$ vs. time | $\|\tilde{\mathbf{d}}\|^2$ vs. samples | $\|\tilde{\mathbf{d}}\|^2$ vs. time

TOFU

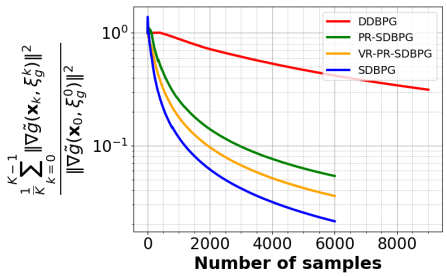


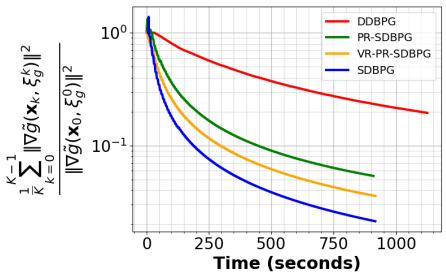


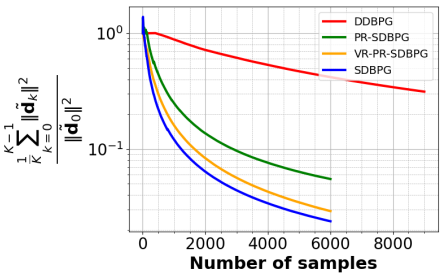


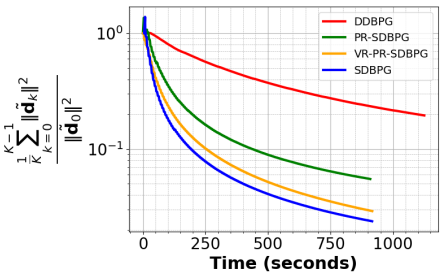


MUSE-News









- Stochastic methods outperform DDBPG in sample efficiency and wall-clock time
- SDBPG fastest, then VR-PR-SDBPG, then PR-SDBPG — matching theory

Thank You!

- Ahmadi, M. M., Cao, J., Mokhtari, A., & Yazdandoost Hamedani, E. (2026). Stochastic Dynamic Barrier Perturbed Gradient Methods for Nonconvex Simple Bilevel Optimization. Advances in Neural Information Processing Systems (NeurIPS 2026).

