A PROXIMAL STOCHASTIC GRADIENT METHOD FOR DOUBLY-REGULARIZED SPECTRAL RISK MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spectral risk minimization (SRM) is an important category of distributionally robust optimization. Recent works in this field elaborate on either distribution shift regularization (DSR) on the spectrum or non-differentiable regularization (NDR) on the parameters. However, few methods can simultaneously handle double regularization. The main difficulty lies in suppressing the bias and variance of the stochastic gradient when double regularization is present. To solve this problem, we develop a novel proximal stochastic gradient method (PSG-SRM) that simultaneously handles double regularization, reduces bias and variance along with iterations, and achieves linear convergence. It has lower computational complexity than two state-of-the-art methods that handle DSR or NDR separately. Experimental results indicate that it achieves competitive performance in both regression and classification tasks, and shows stable performance with respect to randomness.

1 INTRODUCTION

Distributionally robust optimization (DRO) is a general machine learning framework that can be formulated as follows without loss of generality:

$$\min_{w\in\mathbb{R}^d}\max_{\mathscr{Q}\in\mathcal{Q}}\{\mathbb{E}_{X\sim\mathscr{Q}}[\ell(w,X)]-\nu D(\mathscr{Q}\|P_n)\}, \tag{1}$$

where $w\in\mathbb{R}^d$ is a set of parameters with size d, P_n is the empirical distribution of a training set $\{x_i\}_{i=1}^n$ with size n, and $\ell:\mathbb{R}^d\times\mathcal{X}\mapsto\mathbb{R}$ is a loss function. $\nu D(\mathscr{Q}\|P_n)$ is the **distribution shift regularization (DSR)** that measures the price for the test distribution $\mathscr{Q}$ to drift around an uncertainty set of probability measures $\mathcal{Q}$, $D(\mathscr{Q}\|P_n)$ is the divergence (e.g., the χ^2 or K-L divergence) of $\mathscr{Q}$ from P_n on the training set, and $\nu\geqslant 0$ is the shift cost hyperparameter. DRO tries to minimize the expected loss with respect to (w.r.t.) the most adversarial distribution shift in $\mathcal{Q}$ at a price of $\nu D(\mathscr{Q}\|P_n)$. This framework includes both supervised and unsupervised learning according to the data structure of X and the specific form of ℓ.

DRO has inspired a wide range of recent works in learning theory and applications, such as statistical learning (Duchi & Namkoong, 2021), reinforcement learning (Kumar et al., 2020), federated learning (Pillutla et al., 2024), language models (Liu et al., 2021), and robotics (Sharma et al., 2020). In practice, the probability measure $\mathscr{Q}$ can be parameterized by a set of weights $q\in\mathcal{P}:=\left\{p\in\mathbb{R}^n:p\geqslant 0_n,p^\top 1_n=1\right\}$, where $\mathcal{P}$ denotes the n-dimensional simplex (Lai et al., 2018; 2020; Lin et al., 2024a), 0_n and 1_n denote the n-dimensional zero and one vectors, respectively. Accordingly, the empirical distribution P_n is represented by the uniform weights $1_n/n$. Then (1) can be reformulated as the following specific form (Mehta et al., 2024):

$$\min_{w\in\mathbb{R}^d}\mathcal{R}_{\mathcal{P}}(\ell(w)) \text{ for } \mathcal{R}_{\mathcal{P}}(l):=\max_{q\in\mathcal{P}}\left\{\sum_{i=1}^n q_i l_i-\nu D(q\|1_n/n)\right\}, \tag{2}$$

where $\ell(w)=(\ell_1(w),\cdots,\ell_n(w))\in\mathbb{R}^n$ with $\ell_i(w)=\ell(w,x_i)$ being the loss for the i-th training sample. The different font l indicates a table of losses that are maintained while computing $l\approx\ell(w)$ for the current iterate w, which will be explained in Appendix A.3.

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.

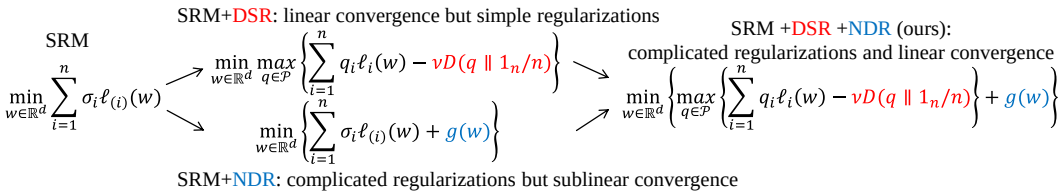

Figure 1: Diagram of doubly-regularized SRM model. It takes advantage of both DSR and NDR that not only facilitates complicated regularizations but also achieves linear convergence.

Spectral risk minimization (SRM) is an important category of DRO that takes the superquantile or extremile (Daouia et al., 2019) as the objective function, such as the conditional value-at-risk (CVaR, Rockafellar & Uryasev 2000; Lin et al. 2024b, an important metric in risk measure), expected shortfall, average top-k loss, and exponential spectral risk measure (ESRM). In general, SRM sorts $\ell_1(w), \cdots, \ell_n(w)$ in an ascending order $\ell_{(1)}(w) \leqslant \cdots \leqslant \ell_{(n)}(w)$, and constructs any vector $\sigma := (\sigma_1, \cdots, \sigma_n) \in \mathcal{P}$ such that $\sigma_1 \leqslant \cdots \leqslant \sigma_n$. Then $\sum_{i=1}^n \sigma_i \ell_{(i)}(w)$ is the maximum of $\sum_{i=1}^n q_i \ell_i(w)$ when q is constrained in the uncertainty set $\mathcal{P}(\sigma) :=$ ConvexHull$\{$permutations of $\sigma\}$. Hence SRM can be specified by using $\mathcal{R}_{\mathcal{P}(\sigma)} =: \mathcal{R}_\sigma$ in (2).

Recently, a new bias and variance reduction approach *PROSPECT* (Mehta et al., 2024) has been proposed for the following strongly convex SRM model:

$$\min_{w\in\mathbb{R}^d} f(w) := \mathcal{R}_\sigma(\ell(w)), \quad \text{each } \ell_i(w) \text{ is } \mu\text{-strongly convex with } \mu > 0. \tag{3}$$

It exploits the μ-strong convexity of f to reduce the bias and variance of the gradient estimation and achieve linear convergence. Another new method *SOREL* (Ge & Jiang, 2025) replaces the DSR in (3) by a μ-strongly convex **non-differentiable regularization (NDR)**:

$$\min_{w\in\mathbb{R}^d} \{\sum_{i=1}^n \sigma_i \ell_{(i)}(w) + g(w)\}, \quad g(w) \text{ is } \mu\text{-strongly convex with } \mu > 0. \tag{4}$$

This strategy enables SOREL to employ the elastic net regularization (Zou & Hastie, 2005) for sparse learning. However, since it drops the DSR, it loses some distributional robustness and only obtains sublinear convergence.

Since NDRs are widely-used to handle complicated tasks, such as out-of-distribution generalization (Lai & Wang, 2024; Wang et al., 2025; Lai et al., 2026a;b), sparse learning (Lin et al., 2024a;b; Lai et al., 2026c), and singularity problems (Lai et al., 2024; 2025), it motivates us to unify NDR and DSR into the following **doubly-regularized SRM model**:

$$\min_{w\in\mathbb{R}^d} \{F(w) := f(w) + g(w) = \max_{q\in\mathcal{P}} \{\sum_{i=1}^n q_i \ell_i(w) - \underbrace{\nu D(q\|1_n/n)}_{DSR}\} + \underbrace{g(w)}_{NDR}\}, \tag{5}$$

where each $\ell_i(w)$ is μ-strongly convex, and $g : \mathbb{R}^d \mapsto \mathbb{R} \cup \{\pm\infty\}$ is a proper, lower semicontinuous, and (not necessarily strongly) convex NDR. The definitions of these terms are provided in Appendix A.1. This model not only enjoys the distributional robustness and linear convergence brought by the DSR, but also allows for more kinds of NDRs than (4). For example, the commonly-used ℓ_1 regularization $g(w) = \lambda\|w\|_1$ with $\lambda \geqslant 0$ cannot be used in (4) due to the lack of strong convexity, but it can be used in (5). The diagram of the proposed doubly-regularized SRM model is shown in Figure 1. Equations (1) to (5) outline a specific development path in this field: (1) DSR $\rightarrow$ (2) SRM+DSR $\rightarrow$ (3) SRM+DSR with strongly convex loss $\rightarrow$ (4) SRM+NDR with strongly convex regularizer $\rightarrow$ (5) SRM+DSR+NDR (ours).

Neat as it is, (5) **is difficult to solve** for several reasons. **1.** Bias and variance reduction algorithms like PROSPECT (Mehta et al., 2024) can handle the DSR but cannot handle NDRs. **2.** Standard proximal (stochastic) gradient methods like SOREL can handle NDRs but cannot handle the DSR. **3.** The convergence results involving the NDR, DSR and bias & variance reduction are difficult to establish, because the bias and variance reduction mechanism only dominates f but not g. To overcome these difficulties, we develop a novel proximal stochastic gradient method for (5), named *PSG-SRM*. **Our main contributions** can be summarized as follows:

1. We skillfully exploit the non-expansiveness of proximal mapping to preserve the bias and variance reduction mechanism for the double regularization in (5), yielding a descent property of PSG-SRM.

2. We develop a variance control scheme for the stochastic gradient, so that this variance can be reduced in the iterative algorithm. Moreover, this variance control scheme accelerates the proximal gradient descent step, which plays a key role in achieving linear convergence.
3. We develop an appropriate Lyapunov function for PSG-SRM to achieve linear convergence, which is better than sublinear convergence that a standard proximal (stochastic) gradient method can achieve (like SOREL). Moreover, the computational complexity of PSG-SRM is even lower than PROSPECT when the size of training set n is also taken into consideration.

To better highlight our contributions, we summarize the requirements, functions, and computational complexities of the three algorithms for SRM in Table 1. PSG-SRM solves the most general problem with the lowest computational complexity among the three.

Table 1: Comparisons between different algorithms for spectral risk minimization. n and ε denote the number of training samples and the convergence tolerance, respectively. The computational complexity of SOREL is non-deterministic w.r.t. n.

Method	Fidelity	DSR	NDR	Computational Complexity
PROSPECT	μ-strongly Convex, G-Lipschitz Continuous, L_ℓ-smooth (All Three)	Available	Unavailable	$\mathcal{O}\left(n \cdot \ln\left(\frac{(C_3'+C_4'n^2+C_1'n)C_2'}{\varepsilon}\right)\right)$
SOREL		Unavailable	μ-strongly Convex	$\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{\varepsilon}}\right)$
PSG-SRM (ours)		Available	Convex	$\mathcal{O}\left(n \cdot \ln\left(\frac{(C_3+\frac{C_1}{n})C_2}{\varepsilon}\right)\right)$

2 Preliminaries and Related Works

Without loss of generality, we review some basic concepts in a way that best fits the context of this paper. A general stochastic gradient descent (SGD) approach aims to construct an estimate v_i for the gradient $\nabla f(w)$ in (3) with only one sample i. This scheme does not need to wait and aggregate the full batch of gradients w.r.t. all the samples, which is suitable for general online learning or stream processing scenarios. Besides, it is empirically faster than the full-batch gradient descent approach at the early stage of the iteration in reducing the loss function. When i is uniformly sampled from $[n] := \{1, 2, \cdots, n\}$ in each update iteration, $v_i \to \nabla f(w)$ holds ideally as the number of iterations increases. To achieve this, it is crucial to control the following expected squared error, which can be decomposed into the bias and variance terms:

$$\mathbb{E}[\|v_i - \nabla f(w)\|_2^2] = \underbrace{\|\mathbb{E}[v_i] - \nabla f(w)\|_2^2}_{\text{bias}} + \underbrace{\mathbb{E}[\|v_i - \mathbb{E}[v_i]\|_2^2]}_{\text{variance}}. \tag{6}$$

A suitable algorithm PROSPECT (Mehta et al., 2024) is developed to simultaneously reduce the above bias and variance along with the iterative process.

2.1 Bias Reduction

The gradient $\nabla f(w) = \nabla_w \mathcal{R}_\sigma(\ell(w))$ is actually computed by the chain rule $(\nabla \ell(w))^\top \nabla_l \mathcal{R}_\sigma(l)$ provided $\ell(w)$ and $\mathcal{R}_\sigma(l)$ satisfy certain smoothness conditions, where $\nabla_l \mathcal{R}_\sigma(l)$ can be computed by Danskin's theorem (Bertsekas, 1999; Mehta et al., 2024):

$$\nabla_l \mathcal{R}_\sigma(l) = \operatorname*{argmax}_{q \in \mathcal{P}(\sigma)} \left\{ \sum_{i=1}^n q_i l_i - \nu D(q \| 1_n/n) \right\}. \tag{7}$$

Details are provided in Appendix A.2. Denote $q^l := \nabla_l \mathcal{R}_\sigma(l)$, since it is actually a set of weights in $\mathcal{P}(\sigma)$. Then $\nabla_w \mathcal{R}_\sigma(\ell(w)) = \sum_{i=1}^n q_i^{\ell(w)} \nabla \ell_i(w)$. In practice, one has to fix $l \approx \ell(w)$ and then solve (7) for q^l via its dual problem, which is illustrated in Appendix A.3. Lemma A2 indicates that the mapping $l \mapsto q^l$ is Lipschitz continuous and thus $\sum_{i=1}^n q_i^l \nabla \ell_i(w) \approx \sum_{i=1}^n q_i^{\ell(w)} \nabla \ell_i(w)$. Hence when i is uniformly sampled from $[n]$, we have $\mathbb{E}[n q_i^l \nabla \ell_i(w)] \approx \nabla_w \mathcal{R}_\sigma(\ell(w))$. This forms an asymptotically unbiased gradient estimate.

2.2 Variance Reduction

As for variance reduction, the control variate method is a widely-used approach in mathematical statistics. Suppose there is an estimator (not necessarily consistent) $\hat{a}$ of $a \in \mathbb{R}^d$, then we can exploit another estimator $\hat{b}$ in the same probability space with a known expectation $\mathbb{E}[\hat{b}] = b \in \mathbb{R}^d$ to construct a new estimator that reduces variance. Specifically, if $\mathbb{E}[(\hat{a} - a)^\top (\hat{b} - b)] > 0$, then the improved

estimator can be $\tilde{a} := \hat{a} - \gamma(\hat{b} - b)$ with $\gamma > 0$ (or $\tilde{a} := \hat{a} + \gamma(\hat{b} - b)$ if $\mathbb{E}[(\hat{a} - a)^\top(\hat{b} - b)] < 0$). This can be seen by

$$\frac{\mathbb{E}[\|\hat{a} - \gamma(\hat{b} - b) - a\|_2^2]}{\mathbb{E}[\|\hat{a} - a\|_2^2]} = 1 - 2\gamma\frac{\mathbb{E}[(\hat{a} - a)^\top(\hat{b} - b)]}{\mathbb{E}[\|\hat{a} - a\|_2^2} + o(\gamma) < 1. \tag{8}$$

In the literature, we let $\hat{a} = nq_i^l\nabla\ell_i(w)$ and $\hat{b} = n\rho_i\nabla\ell_i(z)$. With appropriate designs, $\rho - q^{\ell(w)} \to 0$ and $\nabla\ell(z) - \nabla\ell(w) \to 0$ as the number of iterations increases. Therefore, $\hat{b} - \hat{a} \to 0$. By setting $\gamma = 1$, the stochastic gradient can be set as

$$\hat{a} - \hat{b} + b \approx nq_i^l\nabla\ell_i(w) - n\rho_i\nabla\ell_i(z) + \sum_{j=1}^n \rho_j\nabla\ell_j(z) =: v_i. \tag{9}$$

Suppose $v^{(t)}$ is the sampled stochastic gradient at the t-th iteration, then the variance reduction step aims to asymptotically suppress its variance $\mathrm{Var}(v^{(t)}) \to 0$ as $t \to \infty$. Such a variance reduction scheme accelerates the gradient descent step and leads to linear convergence. This scheme even allows for a differentiable convex regularizer $g(w)$ because such a regularizer can be absorbed in the loss function $\ell(w)$. However, when $g(w)$ is an NDR, this variance reduction scheme fails because the gradient $\nabla g(w)$ does not exist and cannot be absorbed in $\nabla\ell(w)$. Hence the PROSPECT algorithm cannot handle an NDR.

2.3 Primal-dual Approach

As for SRM with only NDR, an intuitive approach is to consider the weights q as dual variables and reformulate (4) as a primal-dual optimization problem (Ge & Jiang, 2025):

$$\begin{aligned} &\min_{w\in\mathbb{R}^d}\max_{q\in\mathcal{P}(\sigma)}\{\mathcal{L}(w,q) := \sum_{i=1}^n q_i\ell_i(w) + g(w)\}, \\ &\begin{cases} q^{(t+1)} = \mathrm{argmax}_{q\in\mathcal{P}(\sigma)}\{\sum_{i=1}^n q_i\ell_i(w^{(t)}) - \frac{1}{\eta^{(t)}}\|q - q^{(t)}\|_2^2\}, \\ w^{(t+1)} = \mathrm{argmin}_{w\in\mathbb{R}^d}\{\sum_{i=1}^n q_i^{(t+1)}\ell_i(w) + g(w) + \frac{1}{\vartheta^{(t)}}\|w - w^{(t)}\|_2^2\}, \end{cases} \end{aligned} \tag{10}$$

where $\eta^{(t)}$ and $\vartheta^{(t)}$ are the dual and primal learning rates, respectively. In both primal and dual updates, proximal terms ($-\frac{1}{\eta^{(t)}}\|q - q^{(t)}\|_2^2$ and $\frac{1}{\vartheta^{(t)}}\|w - w^{(t)}\|_2^2$) are added to stabilize the trajectory of q and facilitate strong convexity on the function $g + \frac{1}{\vartheta^{(t)}}\|\cdot - w^{(t)}\|_2^2$. By carefully setting $\vartheta^{(t)} = \mathcal{O}\left(\frac{1}{t}\right)$, the SOREL algorithm (Ge & Jiang, 2025) can achieve a near optimal sublinear rate of $\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{\varepsilon}}\right)$ for a convergence tolerance of $\varepsilon > 0$. However, since it drops the DSR, it loses some distributional robustness and does not smooth the spectral risk, thus its convergence rate is lower than that of PROSPECT.

3 Methodology

To simultaneously improve distributional robustness, achieve high convergence rate, and allow for complicated regularizations, we propose to solve the doubly-regularized SRM model (5) with both DSR and NDR. In (5), each $\ell_i(w)$ is μ-strongly convex, G-Lipschitz continuous, and L_ℓ-smooth. Note that this can be **satisfied** (Mehta et al., 2024) because the entire **iterative sequence** $\{w^{(t)}\}$ will be restricted to a **compact level set** of the entire objective F (see Appendix A.1).

$$\mu\text{-strongly Convex:} \quad \ell_i(w_2) \geqslant \ell_i(w_1) + (\nabla\ell_i(w_1))^\top(w_2 - w_1) + \frac{\mu}{2}\|w_2 - w_1\|_2^2, \tag{11}$$

$$G\text{-Lipschitz Continuous:} \quad |\ell_i(w_2) - \ell_i(w_1)| \leqslant G\|w_2 - w_1\|_2, \tag{12}$$

$$L_\ell\text{-smooth:} \quad \|\nabla\ell_i(w_2) - \nabla\ell_i(w_1)\|_2 \leqslant L_\ell\|w_2 - w_1\|_2, \quad \forall w_1, w_2 \in \mathbb{R}^d. \tag{13}$$

$g : \mathbb{R}^d \mapsto \mathbb{R} \cup \{\pm\infty\}$ is a proper, lower semicontinuous, and (not necessarily strongly) convex NDR. The detailed explanations of these terms are provided in Appendix A.1.

3.1 Lipschitz Smoothness of f

It directly follows from the G-Lipschitz continuity and smoothness of ℓ that for any $w \in \mathbb{R}^d$,

$$\|\nabla\ell(w)\|_2 \leqslant \|\nabla\ell(w)\|_F = \sqrt{\mathrm{tr}((\nabla\ell(w))^\top\nabla\ell(w))} = \sqrt{\sum_{i=1}^n \|\nabla\ell_i(w)\|_2^2} \leqslant \sqrt{n}G. \tag{14}$$

f is also μ-strongly convex since it is a convex combination of ℓ_i. According to (28) and Lemma A2, $w \mapsto q^{opt}(\ell(w))$ is $L_{q\circ\ell} := \frac{G}{\sqrt{n}\alpha_n \nu}$-Lipschitz continuous (Lemma A3), where α_n denotes the strong convexity parameter associated with the generating function of the divergence $D(q\|1_n/n)$ (see Appendix A.2 for details). Then we can obtain the Lipschitz smoothness of f.

Lemma 1. *For any w,$w' \in \mathbb{R}^d$, $\|\nabla f(w) - \nabla f(w')\|_2 \leqslant L_f \|w - w'\|_2$ for the constant $L_f := M_q L_\ell + \sqrt{n} G L_{q\circ\ell}$, where $M_q := \sup_{w\in\mathbb{R}^d} \|q^{opt}(\ell(w))\|_2$. $M_q < \infty$ because q lies in a compact set $\mathcal{P}(\sigma)$.*

The proof is provided in Appendix A.5.

3.2 Proximal Stochastic Gradient for Spectral Risk Minimization

To better understand the key mechanisms of the proposed PSG-SRM method, we present one iteration of the algorithm along with corresponding explanations in Algorithm 1.

Algorithm 1 PSG-SRM

Input: Initialize: $t = 0$, $\zeta^{(t)} = z^{(t)} = w^{(t)} = 0$. Set learning rate $\eta = \frac{1}{(1+\delta)L_f}, \delta \in (0,1)$, convergence tolerance $\varepsilon > 0$.

1: **for** $t = 0, 1, 2, \cdots, T-1$ **do**
2: Sample two independent indices: $i_t \sim \mathrm{Unif}([n]), j_t \sim \mathrm{Unif}([n])$.
3: Compute stochastic gradient: $v^{(t)} = nq_{i_t}^{(t)} \nabla \ell_{i_t}(w^{(t)}) - (n\rho_{i_t}^{(t)} \nabla \ell_{i_t}(z_{i_t}^{(t)}) - \nabla \bar{\ell}^{(t)})$.
4: Reduce bias: $\zeta_{j_t}^{(t+1)} = w^{(t)}$ and $\zeta_j^{(t+1)} = \zeta_j^{(t)}, \forall j \neq j_t$.
5: Update $q^{(t+1)}$ by (7) with $l^{(t+1)} := \ell(\zeta^{(t+1)})$.
6: Reduce variance: $z_{i_t}^{(t+1)} = w^{(t)}$ and $z_i^{(t+1)} = z_i^{(t)}, \forall i \neq i_t$.
7: $\rho_{i_t}^{(t+1)} = q_{i_t}^{(t)}$ and $\rho_i^{(t+1)} = \rho_i^{(t)}, \forall i \neq i_t$.
8: $\nabla \bar{\ell}^{(t+1)} = \sum_{i=1}^n \rho_i^{(t+1)} \nabla \ell_i(z_i^{(t+1)})$.
9: Proximal stochastic gradient step:
$w^{(t+1)} = \mathrm{prox}_{\eta g}(w^{(t)} - \eta v^{(t)}) := \mathrm{argmin}_{w\in\mathbb{R}^d}\{\frac{1}{2}\|w - (w^{(t)} - \eta v^{(t)})\|_2^2 + \eta g(w)\}$.
10: **end for**

Output: $w^{(T)}$.

Besides, we also present the following three important intermediate variables, although they need not be computed in the iteration.

$$\begin{cases} h^{(t)} = \frac{1}{\eta}(w^{(t)} - w^{(t+1)}). & \text{Update Gap} \\ \Delta^{(t)} = v^{(t)} - \nabla f(w^{(t)}). & \text{Stochastic Gradient Gap} \\ \bar{w}^{(t+1)} = \mathrm{prox}_{\eta g}(w^{(t)} - \eta \sum_{i=1}^n \rho_i^{(t)} \nabla \ell_i(z_i^{(t)})). & \text{Full Proximal Gradient Step} \end{cases} \tag{15}$$

Compared with the PROSPECT algorithm (Mehta et al., 2024), PSG-SRM needs to find a suitable learning rate η according to the Lipschitz constant L_f and implement the proximal stochastic gradient step. Due to this newly-added step, it is nontrivial to establish theoretical results for PSG-SRM, such as convergence properties. **The main difficulty lies in how to suppress the variance of the stochastic gradient gap $\Delta^{(t)}$ in the new scheme.**

To address the above difficulty, we first examine the mean of $\Delta^{(t)}$, and then bound the variance of $\Delta^{(t)}$, by the following lemma.

Lemma 2. *Let w^* be an optimal point and q^* be the corresponding optimal weights. Then*

$$\mathbb{E}_t[\Delta^{(t)}] = 0, \tag{16}$$

$$\mathbb{E}_t[\|\Delta^{(t)}\|_2^2] \leqslant 2Q^{(t)} + 2S^{(t)}, \tag{17}$$

where $Q^{(t)}$ and $S^{(t)}$ are two types of errors:

$$Q^{(t)} := \frac{1}{n}\sum_{i=1}^n \|nq_i^{(t)} \nabla \ell_i(w^{(t)}) - nq_i^* \nabla \ell_i(w^*)\|_2^2,$$

$$S^{(t)} := \frac{1}{n}\sum_{i=1}^n \|n\rho_i^{(t)} \nabla \ell_i(z_i^{(t)}) - nq_i^* \nabla \ell_i(w^*)\|_2^2. \tag{18}$$

Besides, they have the following relationship:

$$\mathbb{E}_t[S^{(t+1)}] := \frac{1}{n}Q^{(t)} + (1-\frac{1}{n})S^{(t)}. \tag{19}$$

The proof is provided in Appendix A.6. (19) is a crucial equation that the conditional expectation of $S^{(t+1)}$ at iteration t can be split into two terms with weights $\frac{1}{n}$ and $(1-\frac{1}{n})$. By this means, part of the error in $S^{(t+1)}$ is transferred to $\frac{1}{n}Q^{(t)}$, which can be reduced by the descent property of PSG-SRM (Lemma 4). As for $(1-\frac{1}{n})S^{(t)}$, it can be reduced with a factor $(1-\frac{1}{n})<1$ at each iteration, which tends to zero in the end. This mechanism finally suppresses the variance of the stochastic gradient gap $\Delta^{(t)}$.

Next, we recall the following lemma for the descent property of a general proximal stochastic gradient step.

Lemma 3 (Descent Property of General Proximal Stochastic Gradient, Xiao & Zhang 2014)**.** *Let $\eta \in (0, \frac{1}{L_f}]$, then for any w,*

$$F(w) \geqslant F(w^{(t+1)}) + \langle h^{(t)}, w - w^{(t)}\rangle + \frac{\eta}{2}\|h^{(t)}\|_2^2 + \frac{\mu}{2}\|w-w^{(t)}\|_2^2 + \langle \Delta^{(t)}, w^{(t+1)} - w\rangle. \tag{20}$$

The proof is provided in Appendix A.7. Then by combining Lemma 2 and Lemma 3, we can obtain the descent property of PSG-SRM. **The key technique is to skillfully exploit the full proximal gradient step $\bar{w}^{(t+1)}$ in (15), which is independent of i_t, to separate the stochastic terms from the deterministic terms in (20).**

Lemma 4 (Descent Property of PSG-SRM)**.**

$$\mathbb{E}_t[\|w^{(t+1)} - w^*\|_2^2] \leqslant (1-\eta\mu)\|w^{(t)} - w^*\|_2^2 - 2\eta[\mathbb{E}_t[F(w^{(t+1)})] - F(w^*)] + 4\eta^2(Q^{(t)} + S^{(t)}), \tag{21}$$

where w^* is a solution to (5).

The proof is provided in Appendix A.8. This lemma indicates that the descent property of PSG-SRM involves both of the argument gap $\|w^{(t+1)} - w^*\|_2$ and the function gap $[\mathbb{E}_t[F(w^{(t+1)})] - F(w^*)]$, which is substantially different from the descent property of PROSPECT (Lemma 11 in Mehta et al. 2024) that only involves the argument gap. The reason is that the strong convexity of f yields the descent of argument gap, while the proximal gradient step yields the descent of the function gap. Besides, we skillfully exploit the non-expansiveness of proximal mapping $\mathrm{prox}_{\eta g}$ when ηg is convex, in order to avoid expansion of the bias and variance.

3.3 Convergence Analysis and Computational Cost

To establish convergence results, it is crucial to develop an appropriate Lyapunov function, which exploits the descent property of both argument gap and function gap, as well as the error splitting in (19). We successfully develop such a Lyapunov function and establish a linear convergence result of PSG-SRM w.r.t. this Lyapunov function.

Theorem 1 (Convergence Theorem)**.** *Consider the following Lyapunov function*

$$V^{(t)} = \|w^{(t)} - w^*\|_2^2 + c_1 \cdot 4\eta^2 S^{(t)}. \tag{22}$$

If the coefficient $c_1 > n$ and the learning rate $\eta \in (0, \frac{\mu}{4L_f(n+c_1)})$, then

$$\begin{cases} \mathbb{E}_t[V^{(t+1)}] \leqslant (1-\tau)V^{(t)} \\ \mathbb{E}[\|w^{(t)} - w^*\|_2^2] \leqslant (1-\tau)^t(1 + 8c_1 n\eta^2 \cdot (\|q^{(0)}\|_2^2 L_\ell^2 + G^2 L_{q\circ\ell}^2))\|w^{(0)} - w^*\|_2^2 \end{cases} \tag{23}$$

holds for any

$$\tau \in (0, \min\{\eta\mu - 4\eta^2 L_f(n+c_1), \frac{1}{n} - \frac{1}{c_1}\}]. \tag{24}$$

It indicates that both the Lyapunov function sequence $\{V^{(t)}\}_{t\geqslant 1}$ and the iterate sequence $\{w^{(t)}\}_{t\geqslant 1}$ converge linearly. By setting the coefficients and hyperparameters as follows:

$$c_1 = 2n, \quad \eta = \frac{\mu}{8L_f(n+c_1)}, \quad \tau = \frac{1}{3n}, \tag{25}$$

the number of iterations of PSG-SRM is $\mathcal{O}\left(n \cdot \ln\left(\frac{(C_3 + \frac{C_1}{n})C_2}{\varepsilon}\right)\right)$ to achieve $\mathbb{E}[\|w^{(t)} - w^*\|_2^2] \leqslant \varepsilon$ for a convergence gap $\varepsilon > 0$, where $C_1, C_2, C_3 \geqslant 0$ are three constants independent of n and ε.

The proof is provided in Appendix A.9. The computational complexity of PSG-SRM is lower than that of PROSPECT $\mathcal{O}\left(n \cdot \ln\left(\frac{(C_3' + C_4' n^2 + C_1' n)C_2'}{\varepsilon}\right)\right)$, considering both n and ε. Note that the standard

proximal (stochastic) gradient method such as SOREL (Ge & Jiang, 2025) or (Xiao & Zhang, 2014) can only achieve a sublinear convergence result $\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{\varepsilon}}\right)$ or $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ due to the non-differentiability of $g(w)$ and the non-contraction of $\mathrm{prox}_{\eta g}$. But in PSG-SRM, the bias and variance reduction scheme is exploited to suppress $V^{(t)}$ in a faster way. Hence PSG-SRM not only reduces the bias and variance of gradient estimation, but also saves computational cost compared with the standard proximal (stochastic) gradient method.

As for the computational complexity of one iteration, the update of $q^{(t+1)}$ by $\zeta^{(t+1)}$ in Algorithm 1 involves a bubble sorting of the losses l and a Pool Adjacent Violators (PAV) algorithm that solves (7). Since there is only one element of l changes in one iteration, the bubble sorting of l requires only $\mathcal{O}(n)$ operations. When $g(w) = \lambda\|w\|_1$ or other similar NDRs (e.g., the elastic net, Zou & Hastie 2005), the proximal stochastic gradient step in Algorithm 1 has a closed-form solution with $\mathcal{O}(d)$ operations. Other steps in Algorithm 1 require either $\mathcal{O}(n)$ or $\mathcal{O}(d)$ operations. Hence the total computational complexity of one iteration is $\mathcal{O}(n+d)$, which is the same as that of PROSPECT, while PROSPECT cannot solve (5). The actual runtime results provided in Appendix 4.4 also show that PSG-SRM achieves faster per-epoch computational speed than PROSPECT in most cases.

4 Experiments

We perform regression and classification tasks to evaluate the effectiveness and robustness of PSG-SRM, compared with PROSPECT, LSVRG (Mehta et al., 2023), and SOREL. More implementation details and experimental results are provided in Appendix B, including spectral risk measures, loss functions, data sets, hyperparameter setting and sensitivity, suboptimality, hyperparamter sensitivity, and more performance comparisons.

4.1 Experimental Settings

Optimization models under spectral risks with different types of regularizations where $g(w)$ is set to be $\lambda\|w\|_1$ (PSG-SRM), $\lambda\|w\|_1 + \frac{1}{2\tau}\|w - w^{(t)}\|_2^2$ (SOREL), $\frac{\lambda}{2}\|w\|_2^2$ (PROSPECT), or 0 (LSVRG) are trained. Note that only PSG-SRM can be used as the optimizer for ℓ_1 regularization, while SOREL uses the elastic net regularization with an accelerated first-order momentum mechanism, which is a strongly-convex regularizer. PROSPECT can be used for the differentiable ℓ_2 regularization, while LSVRG is used for non-regularization (i.e., $g(w) = 0$). Although each competitor can use its own optimization model to perform the tasks in the experiments, only PSG-SRM can solve the doubly-regularized SRM model (5) as a theoretically-sound approach.

Uncertainty sets generated by CVaR, extremile, and ESRM are considered in the experiments (see Appendix B.1). For regression tasks or multi-class classification tasks, the squared loss or the multinomial logistic loss is used for each sample, respectively (see Appendix B.2). The final objective function takes the following form

$$F(w) = \left\{ \max_{q \in \mathcal{P}(\sigma)} \sum_{i=1}^{n} q_i \ell_i(w) - \nu n \|q - 1_n/n\|_2^2 \right\} + g(w),$$

where the χ^2-divergence is instantiated for $D(q\|1_n/n)$. For LSVRG and SOREL, the epoch length is set to the training sample size. The number of training epochs for each data set follows the settings in (Mehta et al., 2024). All the hyperparameters including the shift cost ν, the regularization strength λ, and the learning rate η for LSVRG, PROSPECT, and SOREL are set according to their default criteria, while the hyperparameters for PSG-SRM are set according to Appendix B.4. All the experiments are repeated for 10 times with different random seeds drawn from $\{1, 2, \cdots, 10\}$, then the mean and standard deviation (STD) results are reported.

4.2 Least-Squares Regression

Five benchmarks are used for the regression task under the squared loss: `yacht` ($n = 244$) (Tsanas & Xifara, 2012), `energy` ($n = 614$) (Baressi Šegota et al., 2019), `concrete` ($n = 824$) (Yeh, 2006), `kin8nm` ($n = 6553$) (Akujuobi & Zhang, 2017), and `power` ($n = 7654$) (Tüfekci, 2014). Table 2 presents the average squared losses on the test sets. PSG-SRM achieves the best performance in most cases, especially when the sample sizes are large, like `energy`, `concrete`, `kin8nm`, and `power`. These results indicate that the proposed doubly-regularized SRM model and the PSG-SRM algorithm are effective in these regression tasks. Moreover, PSG-SRM also achieves the most robust performance in nearly all the cases with the lowest STDs. In the situations where different methods

achieve close means (like power), PSG-SRM further achieves the lowest STDs. Hence PSG-SRM is a stable algorithm w.r.t. randomness, which is an advantage as a stochastic method.

Table 2: Performance of different optimizers on regression benchmarks (mean± STD).

Data Set	SRM	PROSPECT	LSVRG	SOREL	PSG-SRM
yacht	CVaR	**0.1989**±1.2e-6	0.2009±1.4e-5	0.1993±1.4e-6	0.2021±**3.7e-16**
	Extremile	**0.1989**±1.0e-6	0.2010±1.2e-5	0.2001±1.1e-7	0.2023±**3.4e-16**
	ESRM	**0.2009**±8.0e-7	0.2056±1.6e-5	0.2012±1.6e-7	0.2039±**4.8e-14**
energy	CVaR	0.0628±9.5e-7	0.0631±2.0e-4	0.0627±1.7e-6	**0.0626**±**2.2e-7**
	Extremile	0.0628±9.4e-7	0.0627±3.1e-6	0.0628±2.4e-7	**0.0626**±**2.2e-7**
	ESRM	0.0628±8.1e-7	0.0627±4.9e-6	0.0627±2.7e-7	**0.0626**±**2.2e-7**
concrete	CVaR	0.2276±7.7e-6	0.2280±1.6e-4	0.2272 ±3.5e-6	**0.2260**±**2.5e-9**
	Extremile	0.2276±7.7e-6	0.2279±5.6e-6	0.2272±2.0e-6	**0.2260**±**2.5e-9**
	ESRM	0.2270±5.2e-6	0.2277±3.4e-6	0.2267 ±2.1e-6	**0.2260**±**2.5e-9**
kin8nm	CVaR	0.2891±8.0e-9	**0.2876**±1.2e-5	0.2877±3.1e-7	0.2877±**1.1e-5**
	Extremile	0.2890±1.2e-7	0.2883 ±1.3e-4	0.2883±**7.3e-9**	**0.2876**±7.8e-9
	ESRM	0.2881±2.8e-9	0.2880 ±4.4e-6	0.2880 ±**7.2e-10**	**0.2876**±7.8e-9
power	CVaR	0.0369±2.3e-10	0.0378±1.3e-4	0.0370±7.5e-7	**0.0369**±**5.6e-11**
	Extremile	0.0369±2.4e-10	0.0369±6.2e-5	0.0369±2.9e-7	**0.0369**±**5.6e-11**
	ESRM	0.0369±2.2e-10	0.0369±7.9e-6	0.0369±7.0e-8	**0.0369**±**5.6e-11**

Following the study of (Williamson & Menon, 2019), we examine how the compared optimizers improve fairness in the regression task. Adult Census (acsincome) is a common benchmark of regression tasks on fairness. It predicts the income of U.S. adults with data collected by the American Community Survey (Ding et al., 2021). The fairness is evaluated by the statistical parity score (SPC, Agarwal et al. 2019), which quantifies discrepancies in outcomes across different values of a protected attribute that is excluded from training. Specifically, denote (X, Y, A) as a random (input, output, attribute) triplet, then a model $\mathcal{G}$ satisfies statistical parity if the conditional distribution of $\mathcal{G}(X)$ over the predicted output under $A = a$ is equal for every value a. In other words, statistical parity measures the maximum gap between these distributions for different values of a, thus a statistical parity score closer to zero indicates higher fairness. In our task, the race and the gender attributes serve as the protected attributes (PA). Results are provided in Table 3. PSG-SRM achieves the lowest statistical parity scores in both spectral risks and both PAs. Specifically, PSG-SRM achieves 70.30% in CVaR and 70.14% in ESRM for the race PA, which are nearly 2% lower than the second best method SOREL. It also achieves 4.52% in CVaR and 4.76% in ESRM for the gender PA, which are nearly 0.2% lower than the second best method PROSPECT. Since SOREL and PROSPECT are two state-of-the-art methods corresponding to the NDR or DSR, respectively, the above results indicate that PSG-SRM achieves good fairness besides competitive regression performance as a method of double regularization.

Table 3: Fairness evaluation of different optimizers on acsincome (mean± STD). A lower value indicates higher fairness. Two protected attributes (PA) are considered: race and gender.

PA	SRM	PROSPECT	LSVRG	SOREL	PSG-SRM
Race	CVaR	0.7540±1.1e-16	0.7340±0.0000	0.7240±0.0000	**0.7030**±1.8e-3
	ESRM	0.7340±0.0000	0.7300±1.1e-16	0.7280±0.0000	**0.7014**±4.4e-3
Gender	CVaR	0.0495±0.0000	0.0510±8.0e-4	0.0514±0.0000	**0.0452**±8.7e-4
	ESRM	0.0494±0.0000	0.0523±0.0000	0.0527±0.0000	**0.0476**±1.9e-4

4.3 Out-of-Distribution Classification

We also consider two classification benchmarks under distribution shifts: amazon and iwildcam. The amazon data set contains product reviews used to predict ratings from 1 to 5, with disjoint sets of reviewers in training and test (Ni et al., 2019). Following the procedure described in (Mehta et al., 2024), we extract review representations by a pretrained BERT model (Devlin et al., 2019) fine-tuned on a held-out subset of the training set for 2 epochs, then adopt the principal component analysis to subsequently reduce dimensionality. The iwildcam data set originates from the iWildCam image

Table 4: Out-of-distribution classification benchmarks on extremile (mean± STD). Performance is evaluated using the worst group test error on `amazon` and the median group test error on `iwildcam`. A lower value indicates better performance.

Data Set	PROSPECT	LSVRG	SOREL	PSG-SRM
`amazon`	0.7791 ±1.1e-16	0.9963±7.3e-3	0.9714±5.2e-2	**0.7674±1.1e-16**
`iwildcam`	0.7532±2.4e-3	0.7483±0.0000	0.7322±0.0000	**0.7085±0.0000**

classification challenge (Beery et al., 2021) and consists of labeled wildlife images captured at various field sites. Distribution shifts arise from changes in camera angles, lighting conditions, locations, and other environmental factors. Deep features for `iwildcam` are extracted by a ResNet-50 model (He et al., 2016) pretrained on ImageNet (Russakovsky et al., 2015) without fine-tuning.

For both `amazon` and `iwildcam`, linear probe classifiers over frozen deep representations are trained with $n = 10000$ and $n = 20000$ examples, respectively. The extremile is used as the spectral risk type because it represents the most adverse situation in the out-of-distribution evaluation. The worst group test errors on `amazon` and the median group test errors on `iwildcam` for the compared optimizers are reported in Table 4. On `amazon`, PSG-SRM achieves the best performance among all the compared optimizers with the lowest worst group test error of 76.74%±1.1e-16, which shows the effectiveness of double regularization with both DSR and NDR. The second best method is PROSPECT, which also has the DSR. However, LSVRG and SOREL deteriorate substantially in performance, reaching test errors higher than 97%. The reason is that LSVRG may fail to converge in some cases, especially for small shift costs. SOREL does not have the DSR, hence it lacks robustness to such an extreme out-of-distribution test scenario. On `iwildcam`, PSG-SRM also shows the best out-of-distribution generalization performance with the lowest median group test error of 70.85%±0.0000, which is nearly 2.5% lower than SOREL and substantially lower than both PROSPECT and LSVRG. The results also indicate that LSVRG and SOREL have better performance in the median group test error than in the worst group test error.

4.4 Runtime Comparison

The total wall-clock times of PROSPECT, SOREL, and PSG-SRM for directly solving the doubly-regularized SRM problem (5) are provided in Table A2. These experiments are conducted on a device with an Intel Core i9-13900K CPU, 64 GB of RAM, and an NVIDIA RTX 4090 GPU. PSG-SRM runs the fastest across different cases, which indicates that it is effective in solving this problem.

5 Conclusion and Discussion

Spectral risk minimization (SRM) is an important methodology that has been widely applied in various fields of machine learning, such as the conditional value-at-risk, expected shortfall, and average top-k loss. There are mainly two kinds of regularizations that can be imposed on SRM to achieve specific purposes. The distribution shift regularization (DSR) on the spectrum improves distributional robustness and the convergence rate, while the non-differentiable regularization (NDR) on the parameters facilitates complicated tasks such as out-of-distribution generalization and sparse learning.

In this study, we propose a doubly-regularized SRM with both DSR and NDR to exploit their advantages. To overcome the difficulties of handling this double regularization, we exploit the non-expansiveness of proximal mapping to develop a variance control scheme for the stochastic gradient, forming a novel proximal stochastic gradient method named PSG-SRM. Moreover, we develop an appropriate Lyapunov function for PSG-SRM to achieve linear convergence, which is superior than the sublinear convergence of a standard proximal (stochastic) gradient method. When taking the sample size into consideration, PSG-SRM also achieves lower computational complexity than two state-of-the-art solvers that handle DSR or NDR separately.

Experimental results show that PSG-SRM achieves competitive performance in regressions, fairness, and out-of-distribution classifications. It also shows stable performance with respect to different random seeds, which is an advantage as a stochastic method. Future works may lie in developing new PSG algorithms for the nonconvex setting of the loss function or the regularization term, or both. Other possible approaches are finding new effective double regularization schemes or developing new distributionally robust optimization models.

REFERENCES

Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 120–129. PMLR, 09–15 Jun 2019.

Uchenna Akujuobi and Xiangliang Zhang. Delve: A dataset-driven scholarly search and analysis system. *SIGKDD Explor. Newsl.*, 19(2):36–46, Nov. 2017. ISSN 1931-0145.

Sandi Baressi Šegota, Nikola Andelić, Jan Kudláček, and Robert Čep. Artificial neural network for predicting values of residuary resistance per unit weight of displacement. *Pomorski zbornik*, 57(1): 9–22, 2019.

Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 2nd edition, 1999.

Abdelaati Daouia, Irène Gijbels, and Gilles Stupfler. Extremiles: A new perspective on asymmetric least squares. *Journal of the American Statistical Association*, 114(527):1366–1381, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6478–6490. Curran Associates, Inc., 2021.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378 – 1406, 2021.

Yuze Ge and Rujun Jiang. Sorel: A stochastic algorithm for spectral risks minimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

Nathaniel Kang, Kibok Lee, and Jongho Im. Calibrated mixup for imbalanced regression on tabular data. *Pattern Recognition*, 172:112610, 2026.

Mahsa Keramati, Lili Meng, and R. Evans. Conr: Contrastive regularizer for deep imbalanced regression. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 26319–26342, 2024.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020.

Zhao-Rong Lai and Weiwen Wang. Invariant risk minimization is a total variation model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 25913–25935. PMLR, 21–27 Jul 2024.

Zhao-Rong Lai, Pei-Yi Yang, Liangda Fang, and Xiaotian Wu. Short-term sparse portfolio optimization based on alternating direction method of multipliers. *Journal of Machine Learning Research*, 19(63):1–28, Oct. 2018.

Zhao-Rong Lai, Liming Tan, Xiaotian Wu, and Liangda Fang. Loss control with rank-one covariance estimate for short-term portfolio optimization. *Journal of Machine Learning Research*, 21(97): 1–37, Jun. 2020.

Zhao-Rong Lai, Xiaotian Wu, Liangda Fang, and Ziliang Chen. A de-singularity subgradient approach for the extended weber location problem. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 4370–4379. International Joint Conferences on Artificial Intelligence Organization, 8 2024.

Zhao-Rong Lai, Xiaotian Wu, Liangda Fang, Ziliang Chen, and Cheng Li. De-singularity subgradient for the q-th-powered ℓ_p-norm weber location problem. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence, AAAI-25*. Association for the Advancement of Artificial Intelligence, 2 2025.

Zhao-Rong Lai, Xiwen Yuan, Ziliang Chen, Liangda Fang, and Yongsen Zheng. A unified total variation framework for membrane potential perturbation dynamic. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Zhao-Rong Lai, Xiwen Yuan, and Jian Weng. Adversarial attack and defense for denoising diffusion sampling. In *The 43rd International Conference on Machine Learning*, 2026b.

Zhao-Rong Lai, Zheng-Sen Zhou, Liangda Fang, Yongsen Zheng, and Ziliang Chen. ℓ_1 latent distance based continuous-time graph representation. In *The Fourteenth International Conference on Learning Representations*, 2026c.

Yizun Lin, Zhao-Rong Lai, and Cheng Li. A globally optimal portfolio for m-sparse sharpe ratio maximization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 17133–17160, 2024a.

Yizun Lin, Yangyu Zhang, Zhao-Rong Lai, and Cheng Li. Autonomous sparse mean-CVaR portfolio optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30440–30456. PMLR, 21–27 Jul 2024b.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 18–24 Jul 2021.

Ronak Mehta, Vincent Roulet, Krishna Pillutla, Lang Liu, and Zaid Harchaoui. Stochastic optimization for spectral risk measures. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10112–10159. PMLR, 25–27 Apr 2023.

Ronak Mehta, Vincent Roulet, Krishna Pillutla, and Zaid Harchaoui. Distributionally robust optimization with bias and variance reduction. In *The Twelfth International Conference on Learning Representations*, 2024.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, September 2018.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93(2): 273–299, 1965.

Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103: 127–152, 2005.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2 edition, 2006.

Krishna Pillutla, Yassine Laguel, Jérôme Malick, and Zaid Harchaoui. Federated learning with superquantile aggregation for heterogeneous data. *Machine Learning*, 113(5):2955–3022, 2024.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Vishnu D Sharma, Maymoonah Toubeh, Lifeng Zhou, and Pratap Tokekar. Risk-aware planning and assignment for ground vehicles using uncertain perception from aerial vehicles. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11763–11769. IEEE, 2020.

Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.

Pınar Tüfekci. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60:126–140, 2014. ISSN 0142-0615.

Yuanchao Wang, Zhao-Rong Lai, and Tianqi Zhong. Out-of-distribution generalization for total variation based invariant risk minimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6786–6797. PMLR, 09–15 Jun 2019.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

I-Cheng Yeh. Analysis of strength of concrete using design of experiments and neural networks. *Journal of Materials in Civil Engineering*, 18(4):597–604, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

A Theoretical Details

A.1 Definitions of Basic Terms

First, we provide some definitions of basic terms used in this paper. Note that we consider the extended real space $\bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ as a general range of function values.

Definition A1 (Convex, Strictly Convex, and Strongly Convex Functions)**.** *A function $h : \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ is called convex if for any $x, y \in \mathbb{R}^n$ and any $\alpha \in (0, 1)$,*

$$h(\alpha x + (1-\alpha)y) \leqslant \alpha h(x) + (1-\alpha)h(y).$$

Furthermore, h is strictly convex if for any $x \neq y$,

$$h(\alpha x + (1-\alpha)y) < \alpha h(x) + (1-\alpha)h(y).$$

In addition, h is μ-strongly convex ($\mu > 0$) if and only if $h - \frac{\mu}{2}\|\cdot\|_2^2$ is convex.

It can be seen that "μ-strong convexity $\Rightarrow$ strict convexity $\Rightarrow$ ordinary convexity" but the converse cannot hold.

Definition A2 (Proper and Lower Semicontinuous Functions)**.** *$h : \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ is a proper function if $\{x \in \mathbb{R}^n : h(x) < +\infty\} \neq \varnothing$ and $-\infty \notin h(\mathbb{R}^n)$. In brief, h has a nonempty domain and never takes $-\infty$ as a function value.*

$h : \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ is lower semicontinuous at a point $x \in \mathbb{R}^n$ if

$$\liminf_{y \neq x,\ y \to x} h(y) := \lim_{\delta \to 0^+} \left(\inf_{y \in B(x;\delta)} h(y) \right) \geqslant h(x),$$

where $B(x;\delta)$ is a neighborhood of x with radius δ. Moreover, if h is lower semicontinuous at every $x \in \mathbb{R}^n$, then h is lower semicontinuous on $\mathbb{R}^n$.

We review a property of convex function, which will be used later.

Lemma A1 (Mehta et al. 2024)**.** *For a convex function $h : \mathbb{R}^n \mapsto \bar{\mathbb{R}}$, if $x_1 \geqslant x_2$ and $y_2 \geqslant y_1$, then*

$$h(y_1 - x_1) + h(y_2 - x_2) \geqslant h(y_2 - x_1) + h(y_1 - x_2). \tag{26}$$

Proof. It can be seen that $(y_2 - x_1)$ and $(y_1 - x_2)$ lie between $(y_1 - x_1)$ and $(y_2 - x_2)$.

$$y_2 - x_1, y_1 - x_2 \in [y_1 - x_1, y_2 - x_2].$$

Hence there exist $\alpha, \beta \in [0, 1]$ such that

$$y_2 - x_1 = \alpha(y_1 - x_1) + (1-\alpha)(y_2 - x_2),$$
$$y_1 - x_2 = \beta(y_1 - x_1) + (1-\beta)(y_2 - x_2).$$

Direct calculation indicates that $\alpha = 1 - \beta$. Since h is convex,

$$h(y_2 - x_1) \leqslant \alpha h(y_1 - x_1) + (1-\alpha)h(y_2 - x_2),$$
$$h(y_1 - x_2) \leqslant (1-\alpha)h(y_1 - x_1) + \alpha h(y_2 - x_2).$$

Summing up both sides leads to (26). □

In the last part of this subsection, we verify that each $\ell_i(w)$ can be simultaneously μ-strongly convex and G-Lipschitz continuous, because the entire iterative sequence $\{w^{(t)}\}$ will be restricted to a compact level set of the entire objective F in (5). Without loss of generality, suppose the optimization starts at $w^{(0)}$. Then a necessary condition of optimizing tractability is the level-boundedness of the entire objective function: $\{w : F(w) \leqslant F(w^{(0)})\}$ is bounded (Rockafellar & Wets, 2009), otherwise $w^{(t)}$ can tend to infinity even if $F(w^{(t)})$ descends. In a d-dimensional real space, this bounded set is also compact. Then based on the descent property Lemma 3, the subsequent level sets $\{w : F(w) \leqslant F(w^{(t)})\}_t$ are all subsets of the initial level set $\{w : F(w) \leqslant F(w^{(0)})\}$, which are all compact. This satisfies the compact set restriction indicated by (Grimmer, 2019), which indicates that $\ell_i(w)$ can be simultaneously μ-strongly convex and G-Lipschitz continuous in the compact set $\{w : F(w) \leqslant F(w^{(0)})\}$. This is sufficient in practical implementation because our algorithm will not go beyond the compact set $\{w : F(w) \leqslant F(w^{(0)})\}$.

A.2 DANSKIN'S THEOREM

Suppose the divergence $D(q\|1_n/n)$ is induced by some generating function $\mathfrak{f}$:

$$D(q\|1_n/n) := \frac{1}{n}\sum_{i=1}^{n}\mathfrak{f}(nq_i). \tag{27}$$

Then as long as $\mathfrak{f}$ is α_n-strongly convex on $[0,n]$, the mapping $q \mapsto \nu D(q\|1_n/n)$ is $(\alpha_n n\nu)$-strongly convex w.r.t. $\|\cdot\|_2$ on $\mathcal{P}(\sigma)$ (Mehta et al., 2024). Moreover, the mapping $q \mapsto \{\sum_{i=1}^{n} q_i l_i - \nu D(q\|1_n/n)\}$ is strongly concave on $\mathcal{P}(\sigma)$ when $\nu > 0$. Because $\mathcal{P}(\sigma)$ is a compact set, $\{\sum_{i=1}^{n} q_i l_i - \nu D(q\|1_n/n)\}$ has a unique maximizer on $\mathcal{P}(\sigma)$:

$$q^{opt}(l) := \operatorname*{argmax}_{q\in\mathcal{P}(\sigma)}\{q^\top l - \nu D(q\|1_n/n)\}. \tag{28}$$

By Danskin's theorem (Bertsekas, 1999; Mehta et al., 2024), this optimal $q^{opt}(l)$ is exactly $\nabla_l \mathcal{R}_\sigma(l)$. Moreover, $\mathcal{R}_\sigma$ is continuously differentiable with this gradient $\nabla_l \mathcal{R}_\sigma(l)$ as $l \in \mathbb{R}^n$.

A.3 SOLVING (7) VIA ITS DUAL PROBLEM

We briefly introduce how (7) can be solved via its dual problem (Mehta et al., 2024) for the convenience of reading and understanding. First, we review an important property (Proposition 6.3.1 of Hiriart-Urruty & Lemaréchal 2004) regarding the conjugate of the sum of two closed convex functions h_1 and h_2:

$$(h_1 + h_2)^*(x) = \inf_{y\in\mathbb{R}^n}\{h_1^*(y) + h_2^*(x-y)\}, \tag{29}$$

where h^* denotes the conjugate function of h:

$$h^*(x) = \sup_{y\in\mathbb{R}^n}\{y^\top x - h(y)\}. \tag{30}$$

Furthermore, if $h_1 + h_2$ is strictly convex, we have

$$\nabla(h_1 + h_2)^*(x) = \operatorname*{argmax}_{u\in\mathbb{R}^n}\{u^\top x - (h_1 + h_2)(u)\}, \tag{31}$$

where the maximizer on the right side is unique. If in addition, $h_1^* + h_2^*$ is strictly convex and h_2^* is differentiable, then again by Danskin's theorem (Bertsekas, 1999; Mehta et al., 2024) we have

$$\nabla(h_1 + h_2)^*(x) = \nabla h_2^*(x - y^*(x)) \;\text{ for }\; y^*(x) = \operatorname*{argmin}_{y\in\mathbb{R}^n}\{h_1^*(y) + h_2^*(x-y)\}. \tag{32}$$

Next, we look into the specific problem to be solved. Define the indicator function of $\mathcal{P}(\sigma)$ as follows:

$$\iota_{\mathcal{P}(\sigma)}(q) := \begin{cases} 0, & \text{if } q \in \mathcal{P}(\sigma); \\ +\infty, & \text{otherwise.} \end{cases} \tag{33}$$

Then its convex conjugate function is

$$\iota^*_{\mathcal{P}(\sigma)}(l) = \max_{q\in\mathcal{P}(\sigma)} q^\top l. \tag{34}$$

Let $h_1(q) = \iota_{\mathcal{P}(\sigma)}(q)$ and $h_2(q) = \Omega(q) := \nu D(q\|1_n/n)$. Assume the divergence generating function $\mathfrak{f}$ in (27) and its conjugate $\mathfrak{f}^*$ are both strictly convex, which is satisfied by some common divergences like χ^2-divergence and KL divergence:

$$\mathfrak{f}_{\chi^2}(x) = x^2 - 1, \quad \mathfrak{f}^*_{\chi^2}(y) = \frac{y^2}{4} + 1, \tag{35}$$

$$\mathfrak{f}_{KL}(x) = x\ln x + \iota_+(x), \quad \mathfrak{f}^*_{KL}(y) = \exp(y-1). \tag{36}$$

Since D in (27) is a sum of $\mathfrak{f}$, it is also strictly convex. Then it follows from (30) that D^* is also strictly convex. Direct calculation yields

$$\sup_{q\in\mathcal{P}(\sigma)}\{q^\top l - \Omega(q)\}$$

$$
\begin{aligned}
&= \sup_{q\in\mathbb{R}^n} \{q^\top l - (\iota_{\mathcal{P}(\sigma)}(q) + \Omega(q))\} \\
&= (\iota_{\mathcal{P}(\sigma)} + \Omega)^*(l) \\
&= \inf_{y\in\mathbb{R}^n} \{\iota^*_{\mathcal{P}(\sigma)}(y) + \Omega^*(l-y)\} \\
&= \inf_{y\in\mathbb{R}^n} \{\max_{q\in\mathcal{P}(\sigma)} q^\top y + \Omega^*(l-y)\} \\
&= \inf_{y\in\mathbb{R}^n} \{\sum_{i=1}^n \sigma_i y_{(i)} + \Omega^*(l-y)\}, \qquad (37)
\end{aligned}
$$

where $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$ are the sorted components of $y \in \mathbb{R}^n$ in the ascending order. On the other hand, combining (31) and (32) yields

$$
\operatorname*{argmax}_{q\in\mathcal{P}(\sigma)} \{q^\top l - \Omega(q)\} = \nabla\Omega^*(l - y^*(l)) \ \text{ for } \ y^*(l) = \operatorname*{argmin}_{y\in\mathbb{R}^n} \{\sum_{i=1}^n \sigma_i y_{(i)} + \Omega^*(l-y)\}. \qquad (38)
$$

(27) indicates that

$$
\Omega(q) = \sum_{i=1}^n \omega(q_i) \ \text{ with } \ \omega(q_i) := \frac{\nu}{n}\mathfrak{f}(nq_i). \qquad (39)
$$

Direct calculation yields

$$
\Omega^*(y) = \sum_{i=1}^n \omega^*(y_i) \ \text{ with } \ \omega^*(y_i) := \frac{\nu}{n}\mathfrak{f}^*(y_i/\nu). \qquad (40)
$$

Lemma A1 indicates that if $l_i \leqslant l_j$ and $y_i \geqslant y_j$, then

$$
\Omega^*(l_i - y_i) + \Omega^*(l_j - y_j) \geqslant \Omega^*(l_i - y_j) + \Omega^*(l_j - y_i). \qquad (41)
$$

Hence if y minimizes $\Omega^*(l-y)$ given l, the ordering of the components of y must follow that of l. That is, for a given permutation $\pi : [n] \mapsto [n]$ such that $l_{\pi(1)} \leqslant \cdots \leqslant l_{\pi(n)}$, we have $y_{\pi(1)} \leqslant \cdots \leqslant y_{\pi(n)}$. On the other hand, the first term of (37) $\sum_{i=1}^n \sigma_i y_{(i)}$ does not depend on the ordering of y. Then (37) can be rewritten as

$$
\begin{aligned}
&\inf_{\substack{y\in\mathbb{R}^n \\ y_{\pi(1)}\leqslant\cdots\leqslant y_{\pi(n)}}} \sum_{i=1}^n \{\sigma_i y_{\pi(i)} + \frac{\nu}{n}\mathfrak{f}^*(\frac{l_{\pi(i)} - y_{\pi(i)}}{\nu})\} \\
= &\inf_{\substack{c\in\mathbb{R}^n \\ c_1\leqslant\cdots\leqslant c_n}} \sum_{i=1}^n \{\sigma_i c_i + \frac{\nu}{n}\mathfrak{f}^*(\frac{l_{\pi(i)} - c_i}{\nu})\} \\
=: &\inf_{\substack{c\in\mathbb{R}^n \\ c_1\leqslant\cdots\leqslant c_n}} \sum_{i=1}^n \mathfrak{g}_i(c_i; l). \qquad (42)
\end{aligned}
$$

Once a solution $c^*(l)$ to the right side of (42) is obtained, the corresponding solution to the left side of (42) can be retrieved by $y^*(l) = c^*_{\pi^{-1}}(l)$, where π^{-1} denotes the inverse permutation of π.

Combining (38), (40), and (42), a solution $q^{opt}(l)$ to (7) can be obtained by

$$
q_i^{opt}(l) = \frac{1}{n}[\mathfrak{f}^*]'(\frac{l_i - c^{opt}_{\mathrm{rank}(i)}(l)}{\nu}), \quad \forall\, i \in [n], \qquad (43)
$$

where $\mathrm{rank}(i)$ denotes the rank of l_i in the sorted form of l in the ascending order. Assuming l is sorted and taking $\mathfrak{f}^*_{\chi^2}$ in (35) as example, we have

$$
c^{opt}(l) = \operatorname*{argmin}_{\substack{c\in\mathbb{R}^n \\ c_1\leqslant\cdots\leqslant c_n}} \sum_{i=1}^n \{\sigma_i c_i + \frac{\nu}{4n}(\frac{l_i - c_i}{\nu})^2\}, \qquad (44)
$$

$$
q_i^{opt}(l) = \frac{1}{2n\nu}(l_i - c_i^{opt}(l)). \qquad (45)
$$

(44) can be solved by letting the derivative w.r.t. each c_i being zero, which yields $c_i^{opt}(l) = l_i + 2n\nu\sigma_i$. Then $q_i^{opt}(l) = \sigma_i$.

A.4 Lipschitz Continuity of $q^{opt}(l)$ and $q^{opt}(\ell(w))$

Lemma A2 (Mehta et al. 2024). *Let $\ell : \mathbb{R}^d \mapsto \mathbb{R}^n$ be differentiable with Jacobian $\nabla\ell(w) \in \mathbb{R}^{n\times d}$ and each ℓ_i be μ-strongly convex. Let $\nu > 0$ and $\mathfrak{f}$ be α_n-strongly convex on $[0, n]$. Then f defined in (3) is differentiable with its gradient being*

$$\nabla f(w) = (\nabla\ell(w))^\top q^{opt}(\ell(w)). \tag{46}$$

Moreover, $l \mapsto q^{opt}(l)$ is $L_q := (\alpha_n n\nu)^{-1}$-Lipschitz continuous.

Proof. From Appendix A.2, $q \mapsto \nu D(q\|1_n/n)$ is $(\alpha_n n\nu)$-strongly convex w.r.t. $\|\cdot\|_2$ on $[0, 1]^n$. Because each ℓ_i is convex w.r.t. w, any $q \in \mathcal{P}(\sigma)$ is nonnegative, and the mapping

$$w \mapsto \max_{q\in\mathcal{P}(\sigma)} \{q^\top\ell(w) - \nu D(q\|1_n/n)\} \tag{47}$$

is a w-point-wise maximum of $\{q^\top\ell(w) - \nu D(q\|1_n/n)\}$ over q, this mapping is convex w.r.t. w. Then by Danskin's theorem in Appendix A.2, f is differentiable with

$$\nabla f(w) = (\nabla\ell(w))^\top q^{opt}(\ell(w)).$$

Moreover, Theorem 1 of (Nesterov, 2005) implies that $l \mapsto q^{opt}(l)$ is Lipschitz continuous with Lipschitz constant being the inverse of the strong convexity constant of $\nu D(q\|1_n/n)$, which is $L_q := (\alpha_n n\nu)^{-1}$. □

Lemma A3 (Mehta et al. 2024). *For any $w, w' \in \mathbb{R}^d$,*

$$\|q^{opt}(\ell(w)) - q^{opt}(\ell(w'))\|_2^2 \leqslant \frac{G^2}{n\alpha_n^2\nu^2}\|w - w'\|_2^2 =: L_{q\circ\ell}^2\|w - w'\|_2^2. \tag{48}$$

Thus the mapping $w \mapsto q^{opt}(\ell(w))$ is $L_{q\circ\ell}$-Lipschitz continuous.

Proof.

$$\begin{aligned}
&\|q^{opt}(\ell(w)) - q^{opt}(\ell(w'))\|_2^2 \\
\leqslant&\frac{1}{n^2\alpha_n^2\nu^2}\|\ell(w) - \ell(w')\|_2^2 &(49)\\
=&\frac{1}{n^2\alpha_n^2\nu^2}\sum_{i=1}^n(\ell_i(w) - \ell_i(w'))^2 \\
\leqslant&\frac{1}{n^2\alpha_n^2\nu^2}\cdot n\cdot G^2\|w - w'\|_2^2 &(50)\\
=&\frac{G^2}{n\alpha_n^2\nu^2}\|w - w'\|_2^2 \\
=:&L_{q\circ\ell}^2\|w - w'\|_2^2. &(51)
\end{aligned}$$

The inequality of (49) is from Lemma A2, and the inequality of (50) is from the G-Lipschitz continuity of ℓ_i. □

A.5 Proof of Lemma 1

Proof. We have $\|q^{opt}(\ell(w))\|_2 \leqslant M_q$ for any $w \in \mathbb{R}^d$ since q lies in a compact set $\mathcal{P}(\sigma)$.

$$\begin{aligned}
&\|\nabla f(w) - \nabla f(w')\|_2 \\
=&\|(\nabla\ell(w))^\top q^{opt}(\ell(w)) - (\nabla\ell(w'))^\top q^{opt}(\ell(w'))\|_2 \\
=&\|(\nabla\ell(w))^\top q^{opt}(\ell(w)) - (\nabla\ell(w'))^\top q^{opt}(\ell(w)) + (\nabla\ell(w'))^\top q^{opt}(\ell(w)) - (\nabla\ell(w'))^\top q^{opt}(\ell(w'))\|_2 \\
\leqslant&\|q^{opt}(\ell(w))\|_2\|\nabla\ell(w) - \nabla\ell(w')\|_2 + \|\nabla\ell(w')\|_2\|q^{opt}(\ell(w)) - q^{opt}(\ell(w'))\|_2 \\
\leqslant&M_qL_\ell\|w - w'\|_2 + \|\nabla\ell(w')\|_2L_{q\circ\ell}\|w - w'\|_2 \\
=&(M_qL_\ell + \|\nabla\ell(w')\|_2L_{q\circ\ell})\|w - w'\|_2 \\
\leqslant&(M_qL_\ell + \sqrt{n}GL_{q\circ\ell})\|w - w'\|_2 \\
=:&L_f\|w - w'\|_2. &(52)
\end{aligned}$$

It indicates that f is Lipschitz smooth, or L_f-smooth to specify the constant L_f. □

A.6 Proof of Lemma 2

Proof. First, we verify that $\Delta^{(t)}$ has zero mean conditioned on time t.

$$\begin{aligned}
&\mathbb{E}_t[\Delta^{(t)}] = \mathbb{E}_t[v^{(t)} - \nabla f(w^{(t)})] \\
&= \mathbb{E}_t[nq_{i_t}^{(t)}\nabla\ell_{i_t}(w^{(t)}) - (n\rho_{i_t}^{(t)}\nabla\ell_{i_t}(z_{i_t}^{(t)}) - \sum_{i=1}^{n}\rho_i^{(t)}\nabla\ell_i(z_i^{(t)})) - \sum_{i=1}^{n}q_i^{(t)}\nabla\ell_i(w^{(t)})] \\
&= \mathbb{E}_t[nq_{i_t}^{(t)}\nabla\ell_{i_t}(w^{(t)}) - n\rho_{i_t}^{(t)}\nabla\ell_{i_t}(z_{i_t}^{(t)})] + \sum_{i=1}^{n}\rho_i^{(t)}\nabla\ell_i(z_i^{(t)}) - \sum_{i=1}^{n}q_i^{(t)}\nabla\ell_i(w^{(t)}) \\
&= \sum_{i=1}^{n}q_i^{(t)}\nabla\ell_i(w^{(t)}) - \sum_{i=1}^{n}\rho_i^{(t)}\nabla\ell_i(z_i^{(t)}) + \sum_{i=1}^{n}\rho_i^{(t)}\nabla\ell_i(z_i^{(t)}) - \sum_{i=1}^{n}q_i^{(t)}\nabla\ell_i(w^{(t)}) \\
&= 0. \qquad (53)
\end{aligned}$$

Next, we turn to bound the variance of $\Delta^{(t)}$ conditioned on time t.

$$\begin{aligned}
&\mathbb{E}_t[\|\Delta^{(t)}\|_2^2] = \mathbb{E}_t[\|v^{(t)} - \nabla f(w^{(t)})\|_2^2] \\
&= \mathbb{E}_t[\|(nq_{i_t}^{(t)}\nabla\ell_{i_t}(w^{(t)}) - \sum_{i=1}^{n}q_i^{(t)}\nabla\ell_i(w^{(t)})) - (n\rho_{i_t}^{(t)}\nabla\ell_{i_t}(z_{i_t}^{(t)}) - \sum_{i=1}^{n}\rho_i^{(t)}\nabla\ell_i(z_i^{(t)}))\|_2^2] \\
&= \mathbb{E}_t[\|(nq_{i_t}^{(t)}\nabla\ell_{i_t}(w^{(t)}) - n\rho_{i_t}^{(t)}\nabla\ell_{i_t}(z_{i_t}^{(t)})) - (\sum_{i=1}^{n}q_i^{(t)}\nabla\ell_i(w^{(t)}) - \sum_{i=1}^{n}\rho_i^{(t)}\nabla\ell_i(z_i^{(t)}))\|_2^2] \qquad (54) \\
&= \mathbb{E}_t[\|nq_{i_t}^{(t)}\nabla\ell_{i_t}(w^{(t)}) - n\rho_{i_t}^{(t)}\nabla\ell_{i_t}(z_{i_t}^{(t)})\|_2^2] - \|\sum_{i=1}^{n}q_i^{(t)}\nabla\ell_i(w^{(t)}) - \sum_{i=1}^{n}\rho_i^{(t)}\nabla\ell_i(z_i^{(t)})\|_2^2 \qquad (55) \\
&\leqslant \mathbb{E}_t[\|nq_{i_t}^{(t)}\nabla\ell_{i_t}(w^{(t)}) - n\rho_{i_t}^{(t)}\nabla\ell_{i_t}(z_{i_t}^{(t)})\|_2^2] \\
&= \mathbb{E}_t[\|(nq_{i_t}^{(t)}\nabla\ell_{i_t}(w^{(t)}) - nq_{i_t}^{*}\nabla\ell_{i_t}(w^{*})) - (n\rho_{i_t}^{(t)}\nabla\ell_{i_t}(z_{i_t}^{(t)}) - nq_{i_t}^{*}\nabla\ell_{i_t}(w^{*}))\|_2^2] \\
&\leqslant 2\mathbb{E}_t[\|nq_{i_t}^{(t)}\nabla\ell_{i_t}(w^{(t)}) - nq_{i_t}^{*}\nabla\ell_{i_t}(w^{*})\|_2^2] + 2\mathbb{E}_t[\|n\rho_{i_t}^{(t)}\nabla\ell_{i_t}(z_{i_t}^{(t)}) - nq_{i_t}^{*}\nabla\ell_{i_t}(w^{*})\|_2^2] \qquad (56) \\
&= \frac{2}{n}\sum_{i=1}^{n}\|nq_i^{(t)}\nabla\ell_i(w^{(t)}) - nq_i^{*}\nabla\ell_i(w^{*})\|_2^2 + \frac{2}{n}\sum_{i=1}^{n}\|n\rho_i^{(t)}\nabla\ell_i(z_i^{(t)}) - nq_i^{*}\nabla\ell_i(w^{*})\|_2^2 \\
&= 2Q^{(t)} + 2S^{(t)}. \qquad (57)
\end{aligned}$$

In (54), we have

$$\mathbb{E}_t[nq_{i_t}^{(t)}\nabla\ell_{i_t}(w^{(t)}) - n\rho_{i_t}^{(t)}\nabla\ell_{i_t}(z_{i_t}^{(t)})] = \sum_{i=1}^{n}q_i^{(t)}\nabla\ell_i(w^{(t)}) - \sum_{i=1}^{n}\rho_i^{(t)}\nabla\ell_i(z_i^{(t)}). \qquad (58)$$

Thus (54) is actually the variance $\mathrm{Var}[nq_{i_t}^{(t)}\nabla\ell_{i_t}(w^{(t)}) - n\rho_{i_t}^{(t)}\nabla\ell_{i_t}(z_{i_t}^{(t)})]$. Then (55) can be derived from (54) by using a property of the variance: $\mathrm{Var}[\|a\|_2^2] = \mathbb{E}[\|a\|_2^2] - \|\mathbb{E}[a]\|_2^2$. The inequality (56) is an arithmetic mean-geometric mean (AM-GM) inequality.

Let $i_t \sim \mathrm{Unif}([n])$ be the random sample index drawn at iteration t, independent of the past history $\mathcal{F}_t$. By definition of the variance term $S^{(t+1)}$:

$$S^{(t+1)} = \frac{1}{n}\sum_{k=1}^{n}\left\|n\rho_k^{(t+1)}\nabla\ell_k(z_k^{(t+1)}) - nq_k^{*}\nabla\ell_k(w^{*})\right\|_2^2. \qquad (59)$$

Taking conditional expectation $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$:

$$\mathbb{E}_t\left[S^{(t+1)}\right] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_t\left[\left\|n\rho_k^{(t+1)}\nabla\ell_k(z_k^{(t+1)}) - nq_k^{*}\nabla\ell_k(w^{*})\right\|_2^2\right]. \qquad (60)$$

In Algorithm 1, for each index $k \in \{1, \ldots, n\}$:

- With probability $\frac{1}{n}$ (when $k = i_t$), $z_k^{(t+1)} = w^{(t)}$ and $\rho_k^{(t+1)} = q_k^{(t)}$.
- With probability $(1 - \frac{1}{n})$ (when $k \neq i_t$), $z_k^{(t+1)} = z_k^{(t)}$ and $\rho_k^{(t+1)} = \rho_k^{(t)}$.

Applying the Law of Total Expectation for each summand k:

$$\begin{aligned}&\mathbb{E}_t\left[\left\|n\rho_k^{(t+1)}\nabla\ell_k(z_k^{(t+1)}) - nq_k^*\nabla\ell_k(w^*)\right\|_2^2\right]\\ =&\frac{1}{n}\left\|nq_k^{(t)}\nabla\ell_k(w^{(t)}) - nq_k^*\nabla\ell_k(w^*)\right\|_2^2 + \left(1-\frac{1}{n}\right)\left\|n\rho_k^{(t)}\nabla\ell_k(z_k^{(t)}) - nq_k^*\nabla\ell_k(w^*)\right\|_2^2. \quad (61)\end{aligned}$$

Summing over $k = 1, \ldots, n$ and considering (18) yield:

$$\mathbb{E}_t[S^{(t+1)}] = \frac{1}{n}Q^{(t)} + (1-\frac{1}{n})S^{(t)}. \quad (62)$$

□

A.7 Proof of Lemma 3

Proof. First, we recall the definition of proximal mapping

$$w^{(t+1)} = \mathrm{prox}_{\eta g}(w^{(t)} - \eta v^{(t)}) = \underset{w\in\mathbb{R}^d}{\mathrm{argmin}}\{\frac{1}{2}\|w - (w^{(t)} - \eta v^{(t)})\|_2^2 + \eta g(w)\}. \quad (63)$$

Since $\eta g(w)$ is convex, Fermat's rule (Proposition 6.a in Moreau 1965) indicates that $w^{(t+1)}$ as a minimizer of (63) satisfies

$$-w^{(t+1)} + (w^{(t)} - \eta v^{(t)}) \in \partial(\eta g(w^{(t+1)})) = \eta\partial g(w^{(t+1)}), \quad (64)$$

where $\partial g(w^{(t+1)})$ denotes the subdifferential of g at $w^{(t+1)}$:

$$\partial g(w^{(t+1)}) := \{y \in \mathbb{R}^d : g(w) \geqslant g(w^{(t+1)}) + \langle y, w - w^{(t+1)}\rangle, \forall w \in \mathbb{R}^d\}. \quad (65)$$

Inserting $y = (w^{(t)} - w^{(t+1)})/\eta - v^{(t)} = h^{(t)} - v^{(t)}$ into (65) and using $y \in \partial g(w^{(t+1)})$ from (64), we have

$$g(w) \geqslant g(w^{(t+1)}) + \langle h^{(t)} - v^{(t)}, w - w^{(t+1)}\rangle, \forall w \in \mathbb{R}^d. \quad (66)$$

On the other hand, since $f(w)$ is μ-strongly convex and L_f-smooth, we have

$$f(w) \geqslant f(w^{(t)}) + \langle\nabla f(w^{(t)}), w - w^{(t)}\rangle + \frac{\mu}{2}\|w - w^{(t)}\|_2^2, \quad (67)$$

$$f(w^{(t)}) \geqslant f(w^{(t+1)}) - \langle\nabla f(w^{(t)}), w^{(t+1)} - w^{(t)}\rangle - \frac{L_f}{2}\|w^{(t+1)} - w^{(t)}\|_2^2. \quad (68)$$

(67) can be deduced from the first-derivative condition of a convex function, while (68) is an important property of L_f-smoothness indicated by Proposition A.24 of (Bertsekas, 1999). Summing up both sides of (66), (67), and (68), and noticing that $f(w) + g(w) = F(w)$ and $f(w^{(t+1)}) + g(w^{(t+1)}) = F(w^{(t+1)})$, we have

$$\begin{aligned}F(w) \geqslant& F(w^{(t+1)}) + \langle\nabla f(w^{(t)}), w - w^{(t+1)}\rangle - \frac{L_f}{2}\|w^{(t+1)} - w^{(t)}\|_2^2\\ &+ \frac{\mu}{2}\|w - w^{(t)}\|_2^2 + \langle h^{(t)} - v^{(t)}, w - w^{(t+1)}\rangle\\ =& F(w^{(t+1)}) - \frac{L_f}{2}\|w^{(t+1)} - w^{(t)}\|_2^2 + \frac{\mu}{2}\|w - w^{(t)}\|_2^2\\ &+ \langle\Delta^{(t)}, w^{(t+1)} - w\rangle + \langle h^{(t)}, w - w^{(t)}\rangle + \langle h^{(t)}, w^{(t)} - w^{(t+1)}\rangle. \quad (69)\end{aligned}$$

Inserting $w^{(t)} - w^{(t+1)} = \eta h^{(t)}$ into (69) yields

$$\begin{aligned}F(w) \geqslant& F(w^{(t+1)}) + (1 - \frac{L_f\eta}{2})\eta\|h^{(t)}\|_2^2 + \frac{\mu}{2}\|w - w^{(t)}\|_2^2\\ &+ \langle\Delta^{(t)}, w^{(t+1)} - w\rangle + \langle h^{(t)}, w - w^{(t)}\rangle. \quad (70)\end{aligned}$$

Since $\eta \leqslant \frac{1}{L_f}$, (70) implies

$$F(w) \geqslant F(w^{(t+1)}) + \frac{\eta}{2}\|h^{(t)}\|_2^2 + \frac{\mu}{2}\|w - w^{(t)}\|_2^2 + \langle\Delta^{(t)}, w^{(t+1)} - w\rangle + \langle h^{(t)}, w - w^{(t)}\rangle. \quad (71)$$

□

A.8 Proof of Lemma 4

Proof. First, we represent the optimality gap of $w^{(t+1)}$ by that of $w^{(t)}$:

$$\begin{aligned}\|w^{(t+1)} - w^*\|_2^2 &= \|w^{(t)} - \eta h^{(t)} - w^*\|_2^2 \\ &= \|w^{(t)} - w^*\|_2^2 - 2\eta\langle h^{(t)}, w^{(t)} - w^*\rangle + \eta^2\|h^{(t)}\|_2^2. \end{aligned} \tag{72}$$

Then we can exploit the descent property (20) in Lemma 3 to suppress the optimality gap:

$$-\langle h^{(t)}, w^{(t)} - w^*\rangle + \frac{\eta}{2}\|h^{(t)}\|_2^2 \leqslant F(w^*) - F(w^{(t+1)}) - \frac{\mu}{2}\|w^* - w^{(t)}\|_2^2 - \langle \Delta^{(t)}, w^{(t+1)} - w^*\rangle. \tag{73}$$

Connecting the left side of (73) to the right side of (72) yields:

$$\|w^{(t+1)} - w^*\|_2^2 \leqslant (1 - \eta\mu)\|w^{(t)} - w^*\|_2^2 + 2\eta[F(w^*) - F(w^{(t+1)})] - 2\eta\langle \Delta^{(t)}, w^{(t+1)} - w^*\rangle. \tag{74}$$

Since the full proximal gradient step

$$\bar{w}^{(t+1)} = \mathrm{prox}_{\eta g}(w^{(t)} - \eta \sum_{i=1}^{n} \rho_i^{(t)} \nabla \ell_i(z_i^{(t)})) \tag{75}$$

is independent of i_t, it can be exploited to separate the stochastic terms from the deterministic terms. Specifically,

$$\begin{aligned} &- 2\eta\langle \Delta^{(t)}, w^{(t+1)} - w^*\rangle \\ =& - 2\eta\langle \Delta^{(t)}, w^{(t+1)} - \bar{w}^{(t+1)}\rangle - 2\eta\langle \Delta^{(t)}, \bar{w}^{(t+1)} - w^*\rangle \\ \leqslant& 2\eta\|\Delta^{(t)}\|_2\|w^{(t+1)} - \bar{w}^{(t+1)}\|_2 - 2\eta\langle \Delta^{(t)}, \bar{w}^{(t+1)} - w^*\rangle \\ \leqslant& 2\eta\|\Delta^{(t)}\|_2\|(w^{(t)} - \eta v^{(t)}) - (w^{(t)} - \eta\nabla f(w^{(t)}))\|_2 - 2\eta\langle \Delta^{(t)}, \bar{w}^{(t+1)} - w^*\rangle \\ =& 2\eta^2\|\Delta^{(t)}\|_2^2 - 2\eta\langle \Delta^{(t)}, \bar{w}^{(t+1)} - w^*\rangle. \end{aligned} \tag{76}$$

Connecting the left side of (76) to the right side of (74) yields

$$\begin{aligned}\|w^{(t+1)} - w^*\|_2^2 \leqslant& (1 - \eta\mu)\|w^{(t)} - w^*\|_2^2 - 2\eta[F(w^{(t+1)}) - F(w^*)] \\ &+ 2\eta^2\|\Delta^{(t)}\|_2^2 - 2\eta\langle \Delta^{(t)}, \bar{w}^{(t+1)} - w^*\rangle. \end{aligned} \tag{77}$$

Taking expectation on both sides of (77) conditioned on i_t yields

$$\begin{aligned}\mathbb{E}_t[\|w^{(t+1)} - w^*\|_2^2] \leqslant& (1 - \eta\mu)\|w^{(t)} - w^*\|_2^2 - 2\eta[\mathbb{E}_t[F(w^{(t+1)})] - F(w^*)] \\ &+ 2\eta^2\mathbb{E}_t[\|\Delta^{(t)}\|_2^2] - 2\eta\mathbb{E}_t[\langle \Delta^{(t)}, \bar{w}^{(t+1)} - w^*\rangle]. \end{aligned} \tag{78}$$

Since $\bar{w}^{(t+1)}$ and w^* are independent of i_t, it follows from (16) that

$$\mathbb{E}_t[\langle \Delta^{(t)}, \bar{w}^{(t+1)} - w^*\rangle] = \langle \mathbb{E}_t[\Delta^{(t)}], \bar{w}^{(t+1)} - w^*\rangle = 0. \tag{79}$$

By this means, the first-order stochastic term of $\Delta^{(t)}$ can be eliminated. Inserting (17) and (79) into (78) yields

$$\mathbb{E}_t[\|w^{(t+1)} - w^*\|_2^2] \leqslant (1 - \eta\mu)\|w^{(t)} - w^*\|_2^2 - 2\eta[\mathbb{E}_t[F(w^{(t+1)})] - F(w^*)] + 4\eta^2(Q^{(t)} + S^{(t)}). \tag{80}$$

□

A.9 Proof of Theorem 1

Proof. First, we verify the linear convergence of the Lyapunov function sequence

$$\mathbb{E}_t[V^{(t+1)}] \leqslant (1 - \tau)V^{(t)}. \tag{81}$$

To do this, we need to bound the error term:

$$\begin{aligned}Q^{(t)} :=&\frac{1}{n}\sum_{i=1}^{n}\|nq_i^{(t)}\nabla\ell_i(w^{(t)}) - nq_i^*\nabla\ell_i(w^*)\|_2^2\\ =&n\|\nabla f(w^{(t)}) - \nabla f(w^*)\|_2^2\\ \leqslant&nL_f\|w^{(t)} - w^*\|_2^2.\end{aligned} \tag{82}$$

Then by direct calculations,

$$\begin{aligned}&\mathbb{E}_t[V^{(t+1)}] = \mathbb{E}_t[\|w^{(t+1)} - w^*\|_2^2] + c_1\cdot 4\eta^2\mathbb{E}_t[S^{(t+1)}]\\ \leqslant&(1-\eta\mu)\|w^{(t)} - w^*\|_2^2 - 2\eta(\mathbb{E}_t[F(w^{(t+1)})] - F(w^*)) + 4\eta^2(Q^{(t)} + S^{(t)}) + c_1\cdot 4\eta^2\mathbb{E}_t[S^{(t+1)}]\\ \leqslant&(1-\eta\mu)\|w^{(t)} - w^*\|_2^2 + 4\eta^2(Q^{(t)} + S^{(t)}) + c_1\cdot 4\eta^2\mathbb{E}_t[S^{(t+1)}]\\ \leqslant&(1-\eta\mu)\|w^{(t)} - w^*\|_2^2 + 4\eta^2(1+\frac{c_1}{n})Q^{(t)} + 4\eta^2(1+c_1-\frac{c_1}{n})S^{(t)}\\ \leqslant&(1-\eta\mu)\|w^{(t)} - w^*\|_2^2 + 4\eta^2L_f(n+c_1)\|w^{(t)} - w^*\|_2^2 + 4\eta^2(1+c_1-\frac{c_1}{n})S^{(t)}\\ \leqslant&(1-\eta\mu + 4\eta^2L_f(n+c_1))\|w^{(t)} - w^*\|_2^2 + 4\eta^2(1+c_1-\frac{c_1}{n})S^{(t)}.\end{aligned} \tag{83}$$

We can properly set the learning step size $\eta \in (0, \frac{\mu}{4L_f(n+c_1)})$ and the coefficient $c_1 > n$, then a suitable τ can be found to satisfy (81):

$$\tau \in (0, \min\{\eta\mu - 4\eta^2L_f(n+c_1), \frac{1}{n} - \frac{1}{c_1}\}]. \tag{84}$$

Second, we verify the linear convergence of the iterate sequence $\{w^{(t)}\}_{t\geqslant 1}$. To do this, we need to bound the initial term:

$$\begin{aligned}S^{(0)} :=&\frac{1}{n}\sum_{i=1}^{n}\|n\rho_i^{(0)}\nabla\ell_i(z_i^{(0)}) - nq_i^*\nabla\ell_i(w^*)\|_2^2\\ =&\frac{1}{n}\sum_{i=1}^{n}\|nq_i^{(0)}\nabla\ell_i(w^{(0)}) - nq_i^*\nabla\ell_i(w^*)\|_2^2\\ =&\frac{1}{n}\sum_{i=1}^{n}\|(nq_i^{(0)}\nabla\ell_i(w^{(0)}) - nq_i^{(0)}\nabla\ell_i(w^*)) + (nq_i^{(0)}\nabla\ell_i(w^*) - nq_i^*\nabla\ell_i(w^*))\|_2^2\\ \leqslant&\frac{2}{n}\sum_{i=1}^{n}\|nq_i^{(0)}(\nabla\ell_i(w^{(0)}) - \nabla\ell_i(w^*))\|_2^2 + \frac{2}{n}\sum_{i=1}^{n}\|n(q_i^{(0)} - q_i^*)\nabla\ell_i(w^*)\|_2^2\\ \leqslant&2n\sum_{i=1}^{n}(q_i^{(0)})^2L_\ell^2\|w^{(0)} - w^*\|_2^2 + 2nG^2\sum_{i=1}^{n}\|q_i^{(0)} - q_i^*\|_2^2\\ =&2n\|q^{(0)}\|_2^2L_\ell^2\|w^{(0)} - w^*\|_2^2 + 2nG^2\|q^{(0)} - q^*\|_2^2\\ \leqslant&2n\|q^{(0)}\|_2^2L_\ell^2\|w^{(0)} - w^*\|_2^2 + 2nG^2L_{q\circ\ell}^2\|w^{(0)} - w^*\|_2^2\\ =&2n(\|q^{(0)}\|_2^2L_\ell^2 + G^2L_{q\circ\ell}^2)\|w^{(0)} - w^*\|_2^2.\end{aligned} \tag{85}$$

On the other hand, since c_1, η, and $S^{(t)} \geqslant 0$, $\|w^{(t)} - w^*\|_2^2 \leqslant V^{(t)}$. Using (81) recursively, we have

$$\begin{aligned}&\mathbb{E}[\|w^{(t)} - w^*\|_2^2]\\ =&\mathbb{E}_0[\|w^{(t)} - w^*\|_2^2]\\ \leqslant&\mathbb{E}_0[V^{(t)}] = \mathbb{E}_0[\mathbb{E}_{t-1}[V^{(t)}]]\\ \leqslant&(1-\tau)\mathbb{E}_0[V^{(t-1)}] \leqslant \cdots\\ \leqslant&(1-\tau)^tV^{(0)}\\ =&(1-\tau)^t(\|w^{(0)} - w^*\|_2^2 + c_1\cdot 4\eta^2S^{(0)})\end{aligned}$$

$$\leqslant (1-\tau)^t(1+8c_1n\eta^2(\|q^{(0)}\|_2^2L_\ell^2+G^2L_{q\circ\ell}^2))\|w^{(0)}-w^*\|_2^2. \tag{86}$$

We can set some specific values for the coefficients and hyperparameters that satisfy linear convergence. For example,

$$c_1=2n,\quad \eta=\frac{\mu}{8L_f(n+c_1)}. \tag{87}$$

Then as long as n is sufficiently large, we can set $\tau=\frac{1}{3n}$. From (51) and (52), we have

$$L_{q\circ\ell}^2=\frac{G^2}{n\alpha_n^2\nu^2},\quad L_f=M_qL_\ell+\sqrt{n}GL_{q\circ\ell}. \tag{88}$$

Inserting (87), (88), and $\tau=\frac{1}{3n}$ into (86) yields

$$\begin{aligned}
&\mathbb{E}[\|w^{(t)}-w^*\|_2^2]\\
\leqslant&(1-\frac{1}{3n})^t(1+\frac{\mu^2}{36(M_qL_\ell+\frac{G^2}{\alpha_n\nu})^2}(\|q^{(0)}\|_2^2L_\ell^2+\frac{G^4}{n\alpha_n^2\nu^2}))\|w^{(0)}-w^*\|_2^2\\
=&(1-\frac{1}{3n})^t(1+\frac{\|q^{(0)}\|_2^2L_\ell^2\mu^2}{36(M_qL_\ell+\frac{G^2}{\alpha_n\nu})^2}+\frac{\mu^2G^4}{36n(M_qL_\ell\alpha_n\nu+G^2)^2})\|w^{(0)}-w^*\|_2^2\\
=:&(1-\frac{1}{3n})^t(C_3+\frac{C_1}{n})C_2,
\end{aligned} \tag{89}$$

where $C_1,C_2,C_3\geqslant 0$ are three constants independent of n.

In order to achieve $\mathbb{E}[\|w^{(t)}-w^*\|_2^2]\leqslant\varepsilon$ for a sufficiently small $\varepsilon>0$, the number of iterations required for PSG-SRM is

$$t=\mathcal{O}\left(\ln\left(\frac{(C_3+\frac{C_1}{n})C_2}{\varepsilon}\right)/\ln\left(\frac{3n}{3n-1}\right)\right), \tag{90}$$

Let $(1-\frac{1}{3n})^t\leqslant\exp(-\frac{t}{3n})$ in (89), then (90) becomes

$$t=\mathcal{O}\left(n\cdot\ln\left(\frac{(C_3+\frac{C_1}{n})C_2}{\varepsilon}\right)\right), \tag{91}$$

which is a lower computational complexity than that of PROSPECT (Mehta et al., 2024) considering both n and ε. The number of iterations for PROSPECT is

$$t=\mathcal{O}\left(n\cdot\ln\left(\frac{(C_3'+C_4'n^2+C_1'n)C_2'}{\varepsilon}\right)\right),\quad C_1',C_2',C_3',C_4'\geqslant 0, \tag{92}$$

which has an extra higher-order term $C_4'n^2$ in the nominator inside the $\ln$ function compared with (91). The reason is that the Lyapunov function of PROSPECT uses $c_1\cdot S^{(0)}$ instead of $c_1\cdot\eta^2S^{(0)}$ and its learning rate η is not divided by any factor of n. This computational complexity may not be further reduced based on the current algorithmic framework of PROSPECT. □

A.10 A Special Case of (2) on Softmax

When the divergence $D(q\|1_n/n)$ is chosen as the KL divergence $D_{KL}(q\|1_n/n)=\sum_{i=1}^n q_i\ln(nq_i)$ and $\mathcal{P}(\sigma)$ is the probability simplex, the inner maximization in (2) yields the closed-form Softmax distribution:

$$q_i^{opt}(l)=\frac{\exp(l_i/\nu)}{\sum_{j=1}^n\exp(l_j/\nu)}. \tag{93}$$

This gives us the following theoretical and empirical implications:

1. For general spectral risks, updating $q^{(t)}$ requires sorting loss components or solving a 1D dual problem via PAV in $\mathcal{O}(n)$ time. In the Softmax special case, $q^{opt}(l)$ is evaluated directly in closed form, which eliminates dual solver overhead.

2. The KL divergence generating function $\mathfrak{f}_{KL}(x) = x \ln x$ is α_n-strongly convex on $[0, n]$ with parameter $\alpha_n = 1/n$. Consequently, $l \mapsto q^{opt}(l)$ is L_q-Lipschitz continuous with $L_q = 1/\nu$.
3. PSG-SRM Advantage: When combining Softmax loss with a non-differentiable parameter regularizer $g(w)$ (such as ℓ_1-norm for sparsity/feature selection):

$$\min_{w\in\mathbb{R}^d} \nu \ln\left(\frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{\ell_i(w)}{\nu}\right)\right) + \lambda\|w\|_1. \tag{94}$$

Existing methods either fail to handle the non-differentiable $g(w)$ (e.g., PROSPECT) or achieve only sublinear rates (e.g., SOREL). But the proposed PSG-SRM solves this Softmax formulation with a guaranteed linear convergence rate $\mathcal{O}(n\ln((C_3 + \frac{C_1}{n})C_2/\varepsilon))$ and $\mathcal{O}(n + d)$ per-iteration complexity.

B Implementation Details and Additional Results for Experiments

B.1 Spectral Risk Measures

Given a spectrum $\sigma := (\sigma_1, \cdots, \sigma_n)$ with $\sigma_1 \leqslant \cdots \leqslant \sigma_n$, the uncertainty set is defined as

$$\mathcal{P}(\sigma) := \text{ConvexHull}\{\text{permutations of } \sigma\}$$

and the corresponding spectral risk measure is

$$\min_{w\in\mathbb{R}^d} \mathcal{R}_{\mathcal{P}(\sigma)}(\ell(w)) \text{ for } \mathcal{R}_{\mathcal{P}(\sigma)}(l) := \max_{q\in\mathcal{P}(\sigma)} \left\{\sum_{i=1}^{n} q_i l_i - \nu D(q\|1_n/n)\right\},$$

where $\ell(w) = (\ell_1(w), \cdots, \ell_n(w)) \in \mathbb{R}^n$ with $\ell_i(w) := \ell(w, x_i, y_i)$ being the loss of the i-th training sample for supervised learning.

We consider three types of spectral risk measures: p-CVaR, b-extremile, and ς-ESRM, defined as follows:

$$\sigma_i = \begin{cases} \frac{1}{np} & \text{if } i \in \{\lceil n(1-p)\rceil, \cdots, n\} \\ 1 - \frac{\lfloor np\rfloor}{np} & \text{if } \lfloor n(1-p)\rfloor < i < \lceil n(1-p)\rceil \end{cases}, \quad p \in (0,1); \tag{p-CVaR}$$

$$\sigma_i = \left(\frac{i}{n}\right)^b - \left(\frac{i-1}{n}\right)^b, \quad b \geq 1; \tag{b-extremile}$$

$$\sigma_i = \frac{e^{\varsigma}\left(e^{\varsigma i/n} - e^{\varsigma(i-1)/n}\right)}{1 - e^{-\varsigma}}, \quad \varsigma > 0. \tag{ς-ESRM}$$

In our experiments, we set $p = 0.5$, $b = 2$, and $\varsigma = 1.0$ for CVaR, extremile, and ESRM, respectively.

B.2 Loss Functions

For regression tasks, given a training sample $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, we use the squared loss

$$\ell_i(w) := \frac{1}{2}\left(y_i - w^\top x_i\right)^2.$$

Note that this setting is based on the evaluating protocols from the corresponding regression tasks and datasets, thus we strictly follow this setting to facilitate a **fair experimental comparison** with other baselines. In fact, it is a **more general problem (i.e., non-strongly-convex)** than what PSG-SRM intends to solve in theory. However, **PSG-SRM still achieves competitive performance in this more general setting** according to the experimental results, which indicates the good extendability of PSG-SRM.

For multi-class classification tasks, given a training sample $(x_i, y_i) \in \mathbb{R}^d \times \{1, 2, \ldots, C\}$, we use the multinomial logistic loss

$$\ell_i(w) := -\log p((x_i, y_i); w), \quad \text{where } p((x_i, y_i); w) := \frac{\exp(w_{y_i}^T x_i)}{\sum_{j=1}^{C}\exp(w_j^T x_i)}.$$

In regression tasks, the parameter w is a d-dimensional column vector. As for multi-class classification tasks, w is a $d \times C$ matrix, where w_j denotes the j-th column of w.

B.3 Data Sets

We provide profiles of the data sets used for the experiments in Table A1. For more details about these data sets, please refer to (Mehta et al., 2024).

Table A1: Profiles of data sets.

Data Set	d	n_{train}	n_{val}	n_{test}	Task
yacht	6	244	31	31	Regression
energy	8	614	77	77	Regression
concrete	8	824	103	103	Regression
kin8nm	8	6553	819	820	Regression
power	4	7654	957	957	Regression
acsincome	202	4000	500	500	Regression
amazon	137	10000	5000	5000	Multi-class Classification
iwildcam	157	20000	5000	5000	Multi-class Classification

B.4 Hyperparameters for PSG-SRM

The regularization parameter λ is selected from the set $\{0.01, 0.1, 1, 10, 100\}$, the learning rate η is selected from the set $\{1e-5, 3e-5, 1e-4, 3e-4, 1e-3, 3e-3, 0.01, 0.03, 0.1\}$, and the shift cost ν is selected from the set $\{0.01, 0.1, 1, 10, 100\}$. The random seed k that determines algorithmic randomness is drawn from $\{1, 2, \cdots, 10\}$. We perform a grid search to determine the optimal λ, η, and ν, then evaluate parameter configurations based on the average validation loss across different random seeds. Specifically, we define the validation loss as

$$\mathcal{L}_k(\lambda, \eta, \nu) = \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} \ell_i(w_k^*(\lambda, \eta, \nu)),$$

where $w_k^*(\lambda, \eta, \nu)$ is learned on the training set by using specific values of λ, η, ν, and k. The final values of λ, η, and ν are chosen such that the average validation loss $\frac{1}{10} \sum_{k=1}^{10} \mathcal{L}_k(\lambda, \eta, \nu)$ is minimized.

B.5 Proximal Operator for SOREL

The implementation of SOREL involves computing the following proximal operator:

$$\text{prox}_{\alpha(g + \frac{1}{2\tau}\|\cdot - u\|^2)}(v) := \underset{w}{\text{argmin}} \left\{ \alpha g(w) + \frac{\alpha}{2\tau}\|w - u\|^2 + \frac{1}{2}\|w - v\|^2 \right\}, \quad \alpha \geqslant 0, \tau > 0,$$

which can be simplified as

$$\text{prox}_{\frac{\tau\alpha}{\alpha+\tau} g} \left(\frac{\alpha}{\alpha + \tau} u + \frac{\tau}{\alpha + \tau} v \right).$$

B.6 Suboptimality Experiment

We follow (Mehta et al., 2024) to conduct the suboptimality experiment. The corresponding objective functions $F(w)$ for PSG-SRM, PROSPECT, and SOREL are given in (5), (3), and (4), respectively. The number of passes k is defined as the number of calls of a single-sample oracle $(\ell_i(w), \nabla \ell_i(w))$ divided by n (i.e., the size of training set). The suboptimality w.r.t. k is calculated by

$$\texttt{Suboptimality}(w_k) = \frac{F(w_k) - F(w^*)}{F(w_0) - F(w^*)},$$

Table A2: Total wall-clock times of PROSPECT, SOREL and PSG-SRM. The runtime is measured in seconds and reported in the form of mean± STD (10 repetitions).

Data Set	n_{train}	SRM	PROSPECT	SOREL	PSG-SRM
yacht	244	CVaR	10.211±0.210	34.720±0.334	**6.364**±0.188
		Extremile	8.264±0.132	35.170±0.616	**6.397**±0.133
		ESRM	11.135±0.166	34.765±0.484	**6.358**±0.113
energy	614	CVaR	26.861±1.006	86.385±1.582	**18.093**± 0.812
		Extremile	34.143±1.812	86.236± 1.783	**17.608**±0.859
		ESRM	33.725±1.805	85.553±1.828	**17.580**±1.188
concrete	824	CVaR	36.908±1.763	115.108±2.757	**25.59**6±0.829
		Extremile	50.368±1.716	114.107±2.969	**25.085**±0.865
		ESRM	50.148±1.591	114.559±3.137	**24.857**±0.890
kin8nm	6553	CVaR	695.67±14.925	454.866±9.981	**333.102**±8.892
		Extremile	891.789±17.002	459.553±9.786	**331.035**±8.705
		ESRM	846.777±20.697	458.549±10.187	**331.815**±9.465
power	7654	CVaR	955.973±18.645	533.611±9.667	**436.830**±7.849
		Extremile	1493.354±14.674	532.237±5.945	**435.958**±4.205
		ESRM	1630.118±24.780	533.489±12.415	**436.108**±13.932

where w^* is an approximated solution solved by LBFGS (Nocedal & Wright, 2006). To make a fair comparison, the same ℓ_1 regularization $g(w) = \lambda\|w\|_1$ is used for both PSG-SRM and SOREL. PROSPECT can be seen as an ablated version of PSG-SRM by setting the NDR as zero. However, PROSPECT uses the differentiable regularization $\lambda\|w\|_2^2/2$ (because it cannot handle $\lambda\|w\|_1$), while PSG-SRM uses the non-differentiable regularization $\lambda\|w\|_1$. Since $\lambda\|w\|_2^2/2$ has better geometrical properties than $\lambda\|w\|_1$, this grants more advantage to PROSPECT against PSG-SRM in the suboptimality experiment. Results in Figure A1 show that PSG-SRM is better than SOREL in most cases, and competitive with PROSPECT in the overall performance. Besides, PSG-SRM is the fastest method that converges to flat plots in most cases.

B.7 Robustness to Hyperparameter Change

Table A3 shows the robustness of PSG-SRM to hyperparameter change on power. For the convenience of demonstration, the regularization parameter is fixed at $\lambda = 1$, while the first row and the first column in each table contain different values of hyperparameters ν (shift cost) and η (learning rate) for grid search, respectively. Results indicate that PSG-SRM achieves stable performance as the learning rate η increases to a sufficiently large level ($0.01 \sim 0.03$) across different values of shift cost ν. Hence PSG-SRM is robust to certain levels of hyperparameter change.

B.8 Comparisons with Stochastic Gradient Descent (SGD)

Standard Gradient Descent (GD) and Stochastic Gradient Descent (SGD) face severe theoretical and practical bottlenecks when applied to doubly-regularized SRM.

1. Full-batch GD: It requires computing gradients over the **entire dataset** at every step ($\mathcal{O}(n \cdot d)$ per update). When an NDR $g(w)$ is present, standard proximal GD only achieves a slow sublinear convergence rate $\mathcal{O}(1/\varepsilon)$ or $\mathcal{O}(1/\sqrt{\varepsilon})$.

2. Standard SGD: Sampling single instances under DSR introduces **non-vanishing variance** and **gradient bias** due to the sorting/permutation step in spectral weights q. Standard SGD fails to converge linearly, and requires decaying learning rates that yield sublinear rates $\mathcal{O}(1/t)$ or $\mathcal{O}(1/\sqrt{t})$.

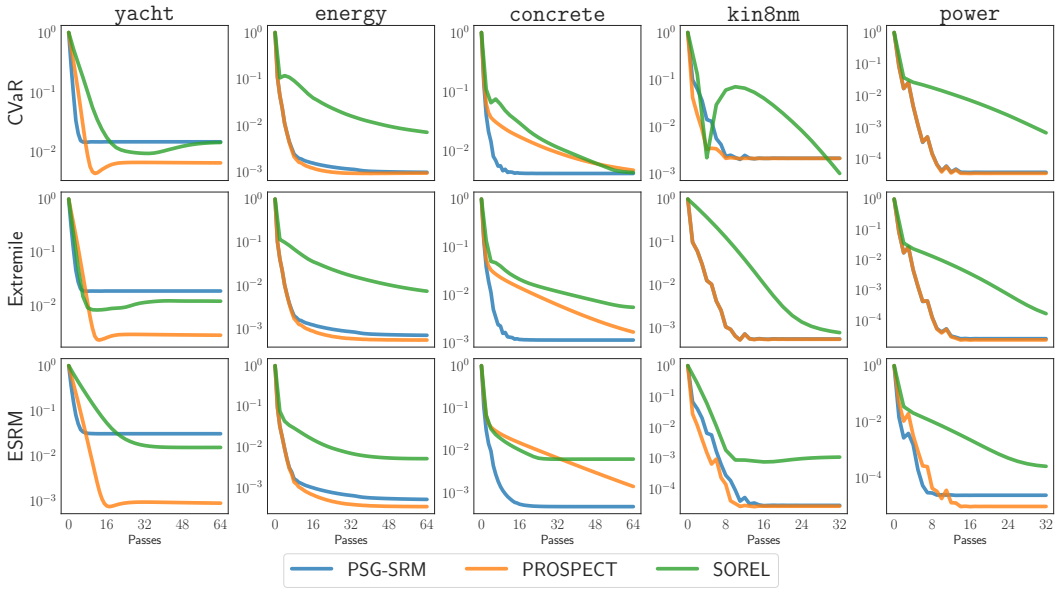

Figure A1: Suboptimality of PSG-SRM, PROSPECT, and SOREL on the regression benchmarks. The y-axis represents the suboptimality and the x-axis represents the number of passes.

Table A3: Performance of PSG-SRM with respect to different hyperparamters.

SRM	η \ ν	0.01	0.1	1	10	100
CVaR	1×10^{-5}	0.4950	0.4950	0.4950	0.4950	0.4950
	3×10^{-5}	0.4950	0.4950	0.4950	0.4950	0.4950
	1×10^{-4}	0.4950	0.4950	0.4950	0.4950	0.4950
	3×10^{-4}	0.1011	0.1116	0.1725	0.2107	0.2166
	1×10^{-3}	0.0480	0.0503	0.0568	0.0586	0.0588
	3×10^{-3}	0.0389	0.0402	0.0434	0.0441	0.0442
	0.01	0.0370	0.0371	0.0374	0.0375	0.0375
	0.03	0.0369	0.0368	0.0369	0.0370	0.0370
Extremile	1×10^{-5}	0.4950	0.4950	0.4950	0.4950	0.4950
	3×10^{-5}	0.4950	0.4950	0.4950	0.4950	0.4950
	1×10^{-4}	0.4950	0.4950	0.4950	0.4950	0.4950
	3×10^{-4}	0.1230	0.1237	0.1725	0.2107	0.2166
	1×10^{-3}	0.0496	0.0505	0.0568	0.0586	0.0588
	3×10^{-3}	0.0395	0.0402	0.0434	0.0441	0.0442
	0.01	0.0370	0.0371	0.0374	0.0375	0.0375
	0.03	0.0369	0.0368	0.0369	0.0370	0.0370
ESRM	1×10^{-5}	0.4950	0.4950	0.4950	0.4950	0.4950
	3×10^{-5}	0.4950	0.4950	0.4950	0.4950	0.4950
	1×10^{-4}	0.4950	0.4950	0.4950	0.4950	0.4950
	3×10^{-4}	0.1566	0.1566	0.1738	0.2107	0.2166
	1×10^{-3}	0.0524	0.0525	0.0567	0.0586	0.0588
	3×10^{-3}	0.0409	0.0410	0.0433	0.0441	0.0442
	0.01	0.0372	0.0372	0.0374	0.0375	0.0375
	0.03	0.0369	0.0369	0.0369	0.0370	0.0370

3. PSG-SRM (Ours): It combines proximal operators for non-differentiable regularizers $g(w)$ with a novel bias and variance control scheme tailored for spectral risk. Practically, this provides two major advantages:

- **Linear Convergence:** It reaches ε-accuracy in $\mathcal{O}(n \ln((C_3 + \frac{C_1}{n})C_2/\varepsilon))$ iterations, and offers exponentially faster convergence than SGD/GD without needing manual step-size decay.
- **Low Per-iteration Cost:** It requires only $\mathcal{O}(n + d)$ operations per iteration (sampling single instances), which makes it drastically cheaper per epoch than full-batch GD while maintaining linear convergence rates.

We additionally include SGD as a baseline for comparison, shown in Table A4. SGD optimizes exactly the same model as PSG-SRM and follows the same hyperparameter selection procedure. We make the following observations:

1. On the `yacht` and `power` data sets in the regression benchmarks (upper table), SGD occasionally achieves a lower average spectral risk than PSG-SRM. Nevertheless, PSG-SRM **outperforms** standard SGD in **most cases**.
2. In the fairness evaluation (middle table) and out-of-distribution classification (lower table) experiments, PSG-SRM **consistently outperforms** SGD.
3. SGD shows **larger standard deviations** than the variance-reduced methods. This observation is consistent with the established theoretical result that, under a constant step size, SGD generally converges only to a neighborhood of the optimum because of the **persistent variance** in stochastic gradient estimates.

Table A4: Performance comparisons of PSG-SRM with SGD (mean $\pm$ STD). **Regression benchmarks (upper):** The test spectral risk are reported. We present the results of our PSG-SRM and SGD for additional comparison. Please refer to the main text of the maunscript for the results of other methods due to space limitations. **Fairness evaluation of different optimizers on acsincome (middle):** A lower value indicates higher fairness. Two protected attributes are considered: race and gender. **Out-of-distribution classification benchmarks on extremile (lower):** Performance is evaluated using the worst group test error on `amazon` and the median group test error on `iwildcam`. A lower value indicates better performance.

SRM	Method	yacht	energy	concrete	kin8nm	power
CVaR	PSG-SRM	0.2021 $\pm$ **3.7e-16**	**0.0626** $\pm$ **2.2e-7**	**0.2260** $\pm$ **2.5e-9**	**0.2877** $\pm$ **1.1e-5**	0.0369 $\pm$ **5.6e-11**
	SGD	**0.2009** $\pm$ 1.2e-3	0.0633 $\pm$ 5.9e-4	0.2278 $\pm$ 1.2e-3	0.2878 $\pm$ 7.9e-5	0.0369 $\pm$ 2.1e-5
Extremile	PSG-SRM	0.2023 $\pm$ **3.4e-16**	**0.0626** $\pm$ **2.2e-7**	**0.2260** $\pm$ **2.5e-9**	**0.2876** $\pm$ **7.8e-9**	0.0369 $\pm$ **5.6e-11**
	SGD	**0.2003** $\pm$ 7.9e-4	0.0632 $\pm$ 4.9e-4	0.2281 $\pm$ 9.5e-4	0.2883 $\pm$ 4.6e-5	**0.0368** $\pm$ 1.2e-5
ESRM	PSG-SRM	0.2039 $\pm$ **4.8e-14**	**0.0626** $\pm$ **2.2e-7**	**0.2260** $\pm$ **2.5e-9**	**0.2876** $\pm$ **7.8e-9**	0.0369 $\pm$ **5.6e-11**
	SGD	**0.2016** $\pm$ 7.6e-4	0.0631 $\pm$ 3.4e-4	0.2267 $\pm$ 2.3e-3	0.2880 $\pm$ 3.9e-5	0.0369 $\pm$ 5.5e-5

SRM	PROSPECT	LSVRG	SOREL	PSG-SRM	SGD
			Protected Attribute: Race		
CVaR	0.7560 ± 0.0000	$0.7300 \pm 1.11 \times 10^{-16}$	$0.7300 \pm 1.11 \times 10^{-16}$	$\mathbf{0.7030} \pm 1.8 \times 10^{-3}$	$0.7250 \pm 1.0 \times 10^{-3}$
ESRM	0.7340 ± 0.0000	$0.7300 \pm 1.11 \times 10^{-16}$	0.7280 ± 0.0000	$\mathbf{0.7014} \pm 4.4 \times 10^{-3}$	$0.7184 \pm 1.7 \times 10^{-3}$
			Protected Attribute: Gender		
CVaR	0.0495 ± 0.0000	$0.0510 \pm 8.0e\text{-}4$	0.0514 ± 0.0000	**0.0452** $\pm$ 8.7e-4	$0.0493 \pm 3.0e\text{-}3$
ESRM	0.0494 ± 0.0000	0.0523 ± 0.0000	0.0527 ± 0.0000	**0.0476** $\pm$ 1.9e-4	$0.0513 \pm 1.3e\text{-}1$

Data Set	PROSPECT	LSVRG	SOREL	PSG-SRM	SGD
`amazon`	$0.7791 \pm 1.11e\text{-}16$	$0.9963 \pm 7.3e\text{-}3$	$0.9714 \pm 5.2e\text{-}2$	$\mathbf{0.7674 \pm 1.11e\text{-}16}$	$0.8070 \pm 5.3e\text{-}3$
`iwildcam`	$0.7532 \pm 2.4e\text{-}3$	0.7483 ± 0.0000	0.7322 ± 0.0000	$\mathbf{0.7085 \pm 0.0000}$	$0.7351 \pm 3.5e\text{-}3$

B.9 Comparisons with χ^2-DRO, Wasserstein DRO, ConR, and CalMix

Classical DRO methods focus on worst-case loss over standard uncertainty balls centered around the empirical distribution. However, they do not accommodate general spectral risk spectrums (e.g., CVaR, extremile, ESRM) alongside non-differentiable parameter regularizers (NDR, like ℓ_1 sparsity) while maintaining linear convergence guarantees. PSG-SRM unifies spectrum-based DSR and non-smooth NDR into a single optimization framework that achieves linear convergence.

We run experimental comparisons between PSG-SRM, Wasserstein DRO (Mohajerin Esfahani & Kuhn, 2018), χ^2-DRO (Ben-Tal et al., 2013), ConR (Keramati et al., 2024), and CalMix (Kang et al., 2026) on the `acsincome` (fairness) and `iwildcam` (out-of-distribution classification) benchmarks.

Note that ConR and CalMix are both designed for imbalanced **regression**, thus they are not applicable to `iwildcam`. Both DRO baselines are implemented using the standard Python package `dro`, running hyperparameter grid searches across uncertainty radii. Both ConR and CalMix are used with their default settings.

- `acsincome`: The ambiguity-set radius for Wasserstein DRO and χ^2 DRO is selected from $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 5 \times 10^{-1}\}$ based on validation-set MSE. We use bias-free OLS as the base model and formulate the resulting **deterministic** convex optimization problems in *CVXPY*.
- `iwildcam`: We implement Wasserstein DRO and χ^2-DRO using the neural-DRO components, with a bias-free 60-class softmax regression model as the linear backbone (trained over 10 independent runs).

Results in Table A5 indicate that PSG-SRM achieves superior fairness trade-offs on `acsincome` and lower group test error on `iwildcam` than other baselines. Besides, PSG-SRM yields consistently lower STDs compared with other baselines.

Table A5: Performance comparisons of PSG-SRM with χ^2-DRO, Wasserstein DRO, ConR, and CalMix.

Fairness Evaluation on `acsincome` (↓)		
Protected Attribute	Race (mean±STD)	Gender (mean±STD)
Wasserstein DRO (deterministic)	0.7140	0.0504
χ^2 DRO (deterministic)	0.7160	0.0561
ConR (10 runs)	0.7474 ± 1.3e-1	0.0474 ± 1.4e-2
CalMix (Nearest Cell) (10 runs)	0.7760 ± 1.0e-1	0.0613 ± 9.8e-3
CalMix (Feature Similarity) (10 runs)	0.8440 ± 3.0e-2	0.0540 ± 3.7e-3
SRM-CVaR (PSG-SRM) (10 runs)	**0.7030** ± 1.8e-3	**0.0452** ± 8.7e-4
SRM-ESRM (PSG-SRM) (10 runs)	**0.7014** ± 4.4e-3	**0.0476** ± 1.9e-4

Out-of-Distribution Classification on `iwildcam` (↓)	
Method	Median Group Test Error (mean±STD)
Wasserstein DRO (10 runs)	0.7117 ± 0.0138
χ^2 DRO (10 runs)	0.7668 ± 0.0243
SRM-Extremile (PSG-SRM) (10 runs)	**0.7085** ± **0.0000**

AI USE STATEMENT

In this work, we used generative AI tools only for grammar correction and minor language edits to improve readability. We have not used generative AI tools for other tasks. We have reviewed all AI-assisted work, and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.