

# A Wasserstein Penalty Framework for Stochastic Optimization

Tao Hu

August 2025

## 1 Introduction

Consider the stochastic optimization problem

$$\min_{\mathbf{B} \in \mathbb{R}^{m \times n}} \mathbb{E}_{\xi \sim \mathbb{P}} f(\mathbf{B}, \xi),$$

where  $\xi \in \Xi$ , write  $f(\mathbf{B}) = \mathbb{E}f(\mathbf{B}, \xi)$ ,  $\nabla f(\mathbf{B}, \xi) = \nabla_{\mathbf{B}} f(\mathbf{B}, \xi) = \frac{\partial f(\mathbf{B}, \xi)}{\partial \mathbf{B}}$ .

Robust optimization has been a very popular topic in the field of optimization.

This article interprets each iteration step in stochastic optimization as a robust decision process.

---

**Algorithm 1** DRO-based Steepest Descent(DROSD)

---

- 1: Initialize  $\mathbf{B}_0 \leftarrow 0$
  - 2: **for**  $t = 0, \dots, T - 1$  **do**
  - 3:   Compute batch gradients  $\mathbf{G}_i \leftarrow \nabla_{\mathbf{B}} f(\mathbf{B}_t, \xi_{t,i})$  for  $i = 1, \dots, N_t$
  - 4:   Compute average  $\bar{\mathbf{G}}_t \leftarrow \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{G}_i$
  - 5:   Obtain  $\mathbf{d}_t$  by solving a DRO subproblem related to the empirical distribution  $\mathbb{P}_{t, N_t} = \sum_{i=1}^{N_t} \delta_{\nabla f(\mathbf{B}_t, \xi_{t,i})}$
  - 6:   Update parameters  $\mathbf{B}_{t+1} \leftarrow \mathbf{B}_t - \mathbf{d}_t$
  - 7: **end for**
  - 8: **return**  $\mathbf{B}_T$
- 

**Assumption 1** (Lipschitz gradient in  $\mathbf{B}$ ). *There exists  $L \geq 0$  such that for all  $\xi$  and all  $\mathbf{B}_1, \mathbf{B}_2 \in \mathbb{R}^{m \times n}$ ,*

$$\|\nabla_{\mathbf{B}} f(\mathbf{B}_1) - \nabla_{\mathbf{B}} f(\mathbf{B}_2)\| \leq L \|\mathbf{B}_1 - \mathbf{B}_2\|_*,$$

where  $\|\cdot\|$  is a chosen matrix norm on  $\mathbb{R}^{m \times n}$ , with dual norm  $\|\cdot\|_*$ . We will take it to be the nuclear norm in this article, and we will explain why it is optimal to use the nuclear norm. Equivalently, for each fixed  $\xi$ , the map  $\mathbf{B} \mapsto f(\mathbf{B}, \xi)$  is  $L$ -smooth with respect to this norm (the constant  $L$  does not depend on  $\xi$ ).

---

**Algorithm 2** Steepest Descent(DROSD)

---

- 1: Initialize  $\mathbf{B}_0 \leftarrow 0, \eta_0 \leftarrow 0$
  - 2: **for**  $t = 0, \dots, T - 1$  **do** **do**
  - 3:   Compute batch gradients  $\mathbf{G}_i \leftarrow \nabla_{\mathbf{B}} f(\mathbf{B}_t, \xi_{t,i})$  for  $i = 1, \dots, N_t$
  - 4:   Compute average  $\overline{\mathbf{G}}_t \leftarrow \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{G}_i$
  - 5:    $\eta_t \leftarrow \beta \eta_{t-1} + (1 - \beta) \|\overline{\mathbf{G}}_t\|$
  - 6:    $\mathbf{d}_t = \eta_t \frac{\overline{\mathbf{G}}_t}{\|\overline{\mathbf{G}}_t\|}$
  - 7:   Update parameters  $\mathbf{B}_{t+1} \leftarrow \mathbf{B}_t - \mathbf{d}_t$
  - 8: **end for**
  - 9: **return**  $\mathbf{B}_T$
- 

---

**Algorithm 3** Steepest Descent

---

- 1: Initialize  $\mathbf{B}_0 \leftarrow 0$
- 2: **for**  $t = 0, \dots, T - 1$  **do** **do**
- 3:   Compute batch gradients  $\mathbf{G}_i \leftarrow \nabla_{\mathbf{B}} f(\mathbf{B}_t, \xi_{t,i})$  for  $i = 1, \dots, N_t$
- 4:   Compute average  $\overline{\mathbf{G}}_t \leftarrow \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{G}_i$
- 5:

$$\mathbf{d}_t = -\max \left\{ 0, \frac{\|\overline{\mathbf{G}}_t\|_{nuc} - C_t}{2L_w} \right\} U_t V_t^\top,$$

where

$U_t \Sigma V_t^\top$  is the thin SVD decomposition of  $\overline{\mathbf{G}}_t$ .

- 6:   Update parameters  $\mathbf{B}_{t+1} \leftarrow \mathbf{B}_t - \mathbf{d}_t$
  - 7: **end for**
  - 8: **return**  $\mathbf{B}_T$
-

## 2 Formulation Using Moment Ambiguity Set

For simplicity of notation, we write  $f(\mathbf{x}, \xi) = f(\mathbf{B}, \xi)$ , where  $\mathbf{x} \in \mathbb{R}^{mn}$  is the vectorization of  $\mathbf{B}$ .

### 2.1 Linear Formulation

Consider the robust optimization problem proposed by Delange and Ye [2]:

$$\begin{aligned} \Psi(\mathbf{x}, \Delta\mathbf{x}, \gamma_1, \gamma_2) = & \text{maximize}_{\mu, f_\xi} \quad \mathbb{E}_{f_\xi}[h(\mathbf{x}, \Delta\mathbf{x}, \xi)] \\ \text{subject to} \quad & \mathbb{E}_{f_\xi}[1] = 1, \quad \mathbb{E}_{f_\xi}[\nabla f(\mathbf{x}, \xi)] = \mu \\ & \mathbb{E}_{f_\xi}[(\nabla f(\mathbf{x}, \xi) - \mu_0)(\nabla f(\mathbf{x}, \xi) - \mu_0)^\top] \preceq \gamma_2 \Sigma_0 \\ & \begin{bmatrix} \Sigma_0 & (\mu - \mu_0) \\ (\mu - \mu_0)^\top & \gamma_1 \end{bmatrix} \succeq 0 \\ & f_\xi(\nabla f(\mathbf{x}, \xi)) \geq 0, \quad \forall \xi \in \mathcal{S}, \end{aligned}$$

where we take  $h(\mathbf{x}, \Delta\mathbf{x}, \xi) = \Delta\mathbf{x}^\top \nabla f(\mathbf{x}, \xi)$ .

**Claim 1.** Take  $\bar{\gamma} = \min\{\gamma_1, \gamma_2\}$ , then

$$\Psi(\mathbf{x}, \Delta\mathbf{x}, \gamma_1, \gamma_2) = \Delta\mathbf{x}^\top \mu_0 + \sqrt{\bar{\gamma}} \sqrt{\Delta\mathbf{x}^\top \Sigma_0 \Delta\mathbf{x}}.$$

*Proof.* In fact, we claim that the maximum is obtained when  $f_\xi$  is supported at a single point.

Since,

$$(\mu - \mu_0)(\mu - \mu_0)^\top + \text{Cov}(\nabla f(\mathbf{x}, \xi)) = \mathbb{E}_{f_\xi}[(\nabla f(\mathbf{x}, \xi) - \mu_0)(\nabla f(\mathbf{x}, \xi) - \mu_0)^\top],$$

we have that

$$(\mu - \mu_0)(\mu - \mu_0)^\top \preceq \mathbb{E}_{f_\xi}[(\nabla f(\mathbf{x}, \xi) - \mu_0)(\nabla f(\mathbf{x}, \xi) - \mu_0)^\top].$$

Therefore,

$$(\mu - \mu_0)(\mu - \mu_0)^\top \preceq \gamma_2 \Sigma_0. \quad (1)$$

If  $\mathbf{y} \in \mathbb{R}^{mn}$  such that  $\Sigma_0 \mathbf{y} = 0$ , then from (4), we know that  $(\mu - \mu_0)^\top \mathbf{y} = 0$ . Therefore,

$$\mu - \mu_0 \in \text{row } \Sigma_0 = \text{col } \Sigma_0.$$

Since

$$\begin{aligned} & \begin{bmatrix} \Sigma_0 & (\mu - \mu_0) \\ (\mu - \mu_0)^\top & \gamma_1 \end{bmatrix} \succeq 0, \\ & \Sigma_0 - \frac{1}{\gamma_1} (\mu - \mu_0)(\mu - \mu_0)^\top \succeq 0. \end{aligned}$$

Therefore,

$$(\mu - \mu_0)(\mu - \mu_0)^\top \preceq \gamma_1 \Sigma_0. \quad (2)$$

Combining (4) and (5), we get that

$$(\mu - \mu_0)(\mu - \mu_0)^T \preceq \bar{\gamma}\Sigma_0. \quad (3)$$

What's more,

$$\mathbb{E}_{f_\xi}[h(\mathbf{x}, \Delta\mathbf{x}, \xi)] = \Delta\mathbf{x}^T \mu.$$

The only restriction for  $\mu$  is (6).

$$\begin{aligned} \Delta\mathbf{x}^T \mu &\leq \Delta\mathbf{x}^T \mu_0 + \Delta\mathbf{x}^T (\mu - \mu_0) \\ &= \Delta\mathbf{x}^T \mu_0 + (\Sigma_0^{1/2} \Delta\mathbf{x})^T \left( \Sigma_0^{\dagger/2} (\mu - \mu_0) \right) \\ &\leq \Delta\mathbf{x}^T \mu_0 + \|\Sigma_0^{1/2} \Delta\mathbf{x}\|_2 \|\Sigma_0^{\dagger/2} (\mu - \mu_0)\|_2 \\ &\leq \Delta\mathbf{x}^T \mu_0 + \bar{\gamma} \|\Sigma_0^{1/2} \Delta\mathbf{x}\|_2 \\ &= \Delta\mathbf{x}^T \mu_0 + \bar{\gamma} \sqrt{\Delta\mathbf{x}^T \Sigma_0 \Delta\mathbf{x}}, \end{aligned}$$

and exact equality holds when

$$\mu = \mu_0 + \frac{\bar{\gamma} \Sigma_0 \Delta\mathbf{x}}{\sqrt{\Delta\mathbf{x}^T \Sigma_0 \Delta\mathbf{x}}}.$$

Therefore,

$$\Psi(\mathbf{x}, \Delta\mathbf{x}, \gamma_1, \gamma_2) = \Delta\mathbf{x}^T \mu_0 + \bar{\gamma} \sqrt{\Delta\mathbf{x}^T \Sigma_0 \Delta\mathbf{x}},$$

and the maximum is obtained when  $f_\xi$  is supported only on one point.  $\square$

## 2.2 RELU Formulation 1

Consider the robust optimization problem proposed by Delange and Ye [2]:

$$\begin{aligned} \Psi(\mathbf{x}, \Delta\mathbf{x}, \gamma_1, \gamma_2) &= \text{maximize}_{\mu, f_\xi} \quad \mathbb{E}_{f_\xi}[h(\mathbf{x}, \Delta\mathbf{x}, \xi)] \\ \text{subject to} \quad &\mathbb{E}_{f_\xi}[1] = 1, \quad \mathbb{E}_{f_\xi}[\nabla f(\mathbf{x}, \xi)] = \mu \\ &\mathbb{E}_{f_\xi}[(\nabla f(\mathbf{x}, \xi) - \mu_0)(\nabla f(\mathbf{x}, \xi) - \mu_0)^T] \preceq \gamma_2 \Sigma_0 \\ &\begin{bmatrix} \Sigma_0 & (\mu - \mu_0) \\ (\mu - \mu_0)^T & \gamma_1 \end{bmatrix} \succeq 0 \\ &f_\xi(\nabla f(\mathbf{x}, \xi)) \geq 0, \quad \forall \xi \in \mathcal{S}, \end{aligned}$$

where we take  $h(\mathbf{x}, \Delta\mathbf{x}, \xi) = -\text{RELU}(-\Delta\mathbf{x}^T \nabla f(\mathbf{x}, \xi))$ .

**Claim 2.** Take  $\bar{\gamma} = \min\{\gamma_1, \gamma_2\}$ , then

$$\Psi(\mathbf{x}, \Delta\mathbf{x}, \gamma_1, \gamma_2) = \min\{0, \Delta\mathbf{x}^T \mu_0 + \sqrt{\bar{\gamma}} \sqrt{\Delta\mathbf{x}^T \Sigma_0 \Delta\mathbf{x}}\}.$$

*Proof.* In fact, we claim that the maximum is obtained when  $f_\xi$  is supported at a single point.

Since,

$$(\mu - \mu_0)(\mu - \mu_0)^T + \text{Cov}(\nabla f(\mathbf{x}, \xi)) = \mathbb{E}_{f_\xi}[(\nabla f(\mathbf{x}, \xi) - \mu_0)(\nabla f(\mathbf{x}, \xi) - \mu_0)^\top],$$

we have that

$$(\mu - \mu_0)(\mu - \mu_0)^T \preceq \mathbb{E}_{f_\xi}[(\nabla f(\mathbf{x}, \xi) - \mu_0)(\nabla f(\mathbf{x}, \xi) - \mu_0)^\top].$$

Therefore,

$$(\mu - \mu_0)(\mu - \mu_0)^T \preceq \gamma_2 \Sigma_0. \quad (4)$$

If  $\mathbf{y} \in \mathbb{R}^{mn}$  such that  $\Sigma_0 \mathbf{y} = 0$ , then from (4), we know that  $(\mu - \mu_0)^T \mathbf{y} = 0$ . Therefore,

$$\mu - \mu_0 \in \text{row } \Sigma_0 = \text{col } \Sigma_0.$$

Since

$$\begin{bmatrix} \Sigma_0 & (\mu - \mu_0) \\ (\mu - \mu_0)^\top & \gamma_1 \end{bmatrix} \succeq 0,$$

$$\Sigma_0 - \frac{1}{\gamma_1}(\mu - \mu_0)(\mu - \mu_0)^T \succeq 0.$$

Therefore,

$$(\mu - \mu_0)(\mu - \mu_0)^T \preceq \gamma_1 \Sigma_0. \quad (5)$$

Combining (4) and (5), we get that

$$(\mu - \mu_0)(\mu - \mu_0)^T \preceq \bar{\gamma} \Sigma_0. \quad (6)$$

What's more, from the concavity of  $-\text{RELU}(-\lambda)$ ,

$$\mathbb{E}_{f_\xi}[h(\mathbf{x}, \Delta \mathbf{x}, \xi)] \leq \min\{0, \mathbb{E}_{f_\xi}(\Delta \mathbf{x}^T \nabla f(\mathbf{x}, \xi))\} = \min\{0, \Delta \mathbf{x}^T \mu\}.$$

The only restriction for  $\mu$  is (6).

$$\begin{aligned} \Delta \mathbf{x}^T \mu &\leq \Delta \mathbf{x}^T \mu_0 + \Delta \mathbf{x}^T (\mu - \mu_0) \\ &= \Delta \mathbf{x}^T \mu_0 + (\Sigma_0^{1/2} \Delta \mathbf{x})^T \left( \Sigma_0^{\dagger/2} (\mu - \mu_0) \right) \\ &\leq \Delta \mathbf{x}^T \mu_0 + \|\Sigma_0^{1/2} \Delta \mathbf{x}\|_2 \|\Sigma_0^{\dagger/2} (\mu - \mu_0)\|_2 \\ &\leq \Delta \mathbf{x}^T \mu_0 + \bar{\gamma} \|\Sigma_0^{1/2} \Delta \mathbf{x}\|_2 \\ &= \Delta \mathbf{x}^T \mu_0 + \bar{\gamma} \sqrt{\Delta \mathbf{x}^T \Sigma_0 \Delta \mathbf{x}}, \end{aligned}$$

and exact equality holds when

$$\mu = \mu_0 + \frac{\bar{\gamma} \Sigma_0 \Delta \mathbf{x}}{\sqrt{\Delta \mathbf{x}^T \Sigma_0 \Delta \mathbf{x}}}.$$

Therefore,

$$\Psi(\mathbf{x}, \Delta \mathbf{x}, \gamma_1, \gamma_2) = \Delta \mathbf{x}^T \mu_0 + \bar{\gamma} \sqrt{\Delta \mathbf{x}^T \Sigma_0 \Delta \mathbf{x}},$$

and the maximum is obtained when  $f_\xi$  is supported only at one point.  $\square$

### 3 Formulation Using a Second Order Wasserstein Distance Regularization

Fixing an iteration  $t$  in Algorithm 1, we solve the following DRO subproblem to obtain  $d_t$ :

$$\mathbf{d}_t = \arg \min_{\mathbf{d} \in \mathbb{R}^{m \times n}} \sup_{\substack{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^{m \times n}) \\ \mathbb{Q}(\mathbb{R}^{m \times n})=1}} \left\{ \mathbb{E}_{G \sim \mathbb{Q}}[\langle \mathbf{d}, G \rangle] - \frac{1}{2\kappa} W_1^2(\mathbb{P}_{t, N_t}, \mathbb{Q}) \right\}, \quad (\text{P2})$$

where  $\langle \mathbf{d}, \mathbf{G} \rangle = \text{trace}(\mathbf{d}^T \mathbf{G})$  and type-p Wasserstein distance  $W_p(\mathbb{Q}_1, \mathbb{Q}_2)$  is defined as follows:

**Definition 1** (Wasserstein Distance).

$$W_p(\mathbb{Q}_1, \mathbb{Q}_2) = \sqrt[p]{\inf_{\pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2)} \int_{\mathbb{R}^m \times \mathbb{R}^m} \|\xi - \xi'\|^p \pi(d\xi, d\xi')},$$

where  $\Pi(\mathbb{Q}_1, \mathbb{Q}_2)$  is the collection of distributions on  $\mathbb{R}^m \times \mathbb{R}^m$  whose marginal distribution with respect to the first  $m$  components is  $\mathbb{Q}_1$  and the marginal distribution with respect to the last  $m$  components is  $\mathbb{Q}_2$ .

In the later proofs, we will make use of an important equivalent characterization of Wasserstein distance which is stated below:

**Theorem 1** (Kantorovich-Rubinstein[3]).

$$W_1(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{f \text{ is 1-Lipschitz continuous}} \left( \int_{\mathbb{R}^{m \times n}} f(\xi) \mathbb{Q}_1(d\xi) - \int_{\mathbb{R}^{m \times n}} f(\xi') \mathbb{Q}_2(d\xi') \right).$$

**Theorem 2.** Problem (P2) is equivalent to

$$\begin{aligned} \mathbf{d}_t &= \arg \min_{\mathbf{d} \in \mathbb{R}^{m \times n}} \left\{ \mathbb{E}_{\mathbf{G} \sim \mathbb{P}_{t, N_t}}[\langle \mathbf{d}, \mathbf{G} \rangle] + \frac{1}{2} \kappa \|\mathbf{d}\|_*^2 \right\} \\ &= \arg \min_{\mathbf{d} \in \mathbb{R}^{m \times n}} \left( \langle \mathbf{d}, \overline{\mathbf{G}}_t \rangle + \frac{1}{2} \kappa \|\mathbf{d}\|_*^2 \right), \end{aligned} \quad (\text{D2})$$

where

$$\overline{\mathbf{G}}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \nabla f(\mathbf{B}_t, \xi_{t,i}).$$

**Theorem 3.** If  $\|\cdot\| = \|\cdot\|_{nuc}$ , then the optimal solution of (D2) is

$$\mathbf{d}_t = -\frac{1}{\kappa} \|\overline{\mathbf{G}}_t\|_{nuc} U_t V_t^\top,$$

where

$U_t \Sigma V_t^\top$  is the thin SVD decomposition of  $\overline{\mathbf{G}}_t$ .

## 4 Formulation Using $p$ -th Order Wasserstein Distance Regularization( $p \in (1, +\infty)$ )

In order to generalize the results in the last section, we require that

$$\mathbf{d}_t = \arg \min_{\mathbf{d} \in \mathbb{R}^{m \times n}} \sup_{\substack{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^{m \times n}) \\ \mathbb{Q}(\mathbb{R}^{m \times n})=1}} \left\{ \mathbb{E}_{G \sim \mathbb{Q}}[\langle \mathbf{d}, G \rangle] - \frac{1}{\kappa p} W_1^p(\mathbb{P}_N, \mathbb{Q}) \right\}, \quad (\text{Pp})$$

where  $p \in (1, +\infty)$ .

**Theorem 4.** *Problem (Pp) is equivalent to*

$$\begin{aligned} \mathbf{d}_t &= \arg \min_{\mathbf{d} \in \mathbb{R}^{m \times n}} \left\{ \mathbb{E}_{G \sim \mathbb{P}_{t, N_t}}[\langle \mathbf{d}, \mathbf{G} \rangle] + \frac{1}{q} \kappa \|\mathbf{d}\|_*^q \right\} \\ &= \arg \min_{\mathbf{d} \in \mathbb{R}^{m \times n}} \left( \langle \mathbf{d}, \overline{\mathbf{G}}_t \rangle + \frac{1}{q} \kappa \|\mathbf{d}\|_*^q \right), \end{aligned} \quad (\text{Dp})$$

where  $\frac{1}{p} + \frac{1}{q} = 1$  and

$$\overline{\mathbf{G}}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \nabla f(\mathbf{B}_t, \xi_{t,i}).$$

**Theorem 5.** *If  $\|\cdot\| = \|\cdot\|_{nuc}$ , the closed-form solution of (Dp) is*

$$\mathbf{d}_t = - \left( \frac{\|\overline{\mathbf{G}}_t\|_{nuc}}{\kappa} \right)^{\frac{1}{q-1}} U_t V_t^\top,$$

where

$$U_t \Sigma V_t^\top \text{ is the thin SVD decomposition of } \overline{\mathbf{G}}_t.$$

## 5 Convergence of MUON beyond Uniform Lipschitz Gradient

**Assumption 2** (Wasserstein Lipschitz Gradient). *Denote  $\mathbb{P}_{\mathbf{B}}$  as the distribution of  $\nabla f(\mathbf{B}, \xi)$ . We assume that*

$$W_1(\mathbb{P}_{\mathbf{B}}, \mathbb{P}_{\mathbf{B}'}) \leq L_w \|\mathbf{B} - \mathbf{B}'\|_*, \forall \mathbf{B}, \mathbf{B}'.$$

where  $W_1$  is the type-1 Wasserstein distance.

There exists  $C > 0$  such that  $W_1(\mathbb{P}_{t,N}, \mathbb{P}_{\mathbf{B}_t}) \leq C$  with high probability. From Assumption 2, we know that  $W_1(\mathbb{P}_{\mathbf{B}_t}, \mathbb{P}_{\mathbf{B}}) \leq L_w \|\mathbf{B}_t - \mathbf{B}\|_*, \forall \mathbf{B} \in \mathbb{R}^{m \times n}$ . Thus,  $\forall \mathbf{B} \in \mathbb{R}^{m \times n}$ , we have that, with high probability,

$$W_1(\mathbb{P}_{t,N}, \mathbb{P}_{\mathbf{B}}) \leq W_1(\mathbb{P}_{t,N}, \mathbb{P}_{\mathbf{B}_t}) + W_1(\mathbb{P}_{\mathbf{B}_t}, \mathbb{P}_{\mathbf{B}}) = C + L_w \|\mathbf{B}_t - \mathbf{B}\|_*.$$

We turn to solve the following subproblem:

$$\mathbf{d}_t = \arg \min_{\mathbf{d} \in \mathbb{R}^{m \times n}} \sup_{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^{m \times n}), \mathbb{Q}(\mathbb{R}^{m \times n})=1, W(\mathbb{P}_t, N_t, \mathbb{Q}) \leq C + L \|\mathbf{d}\|_*} \mathbb{E}_{\mathbf{G} \sim \mathbb{Q}}[\langle \mathbf{d}, \mathbf{G} \rangle]. \quad (\text{P})$$

**Theorem 6.** *Problem (P) is equivalent to*

$$\mathbf{d}_t = \arg \min_{\mathbf{d} \in \mathbb{R}^{m \times n}} (\langle \mathbf{d}, \overline{\mathbf{G}}_t \rangle + C \|\mathbf{d}\|_* + L_w \|\mathbf{d}\|_*^2), \quad (\text{D})$$

where

$$\overline{\mathbf{G}}_t = \frac{1}{N} \sum_{i=1}^N \nabla f(\mathbf{B}_t, \xi_{t,i}).$$

**Theorem 7.** *If  $\|\cdot\| = \|\cdot\|_{\text{nuc}}$ , the closed-form solution of (D) is*

$$\mathbf{d}_t = -\max \left\{ 0, \frac{\|\overline{\mathbf{G}}_t\|_{\text{nuc}} - C}{2L_w} \right\} U_t V_t^\top,$$

where

$$U_t \Sigma V_t^\top \text{ is the thin SVD decomposition of } \overline{\mathbf{G}}_t.$$

## 5.1 Convergence Rate Analysis Beyond Nuclear Norm

**Assumption 3** (Light-tail Distribution). *There exists a uniform constant  $a > 1$  and  $A > 0$  such that*

$$\mathbb{E}_{\mathbf{G} \sim \mathbb{P}_{\mathbf{B}}}[\exp(\|\mathbf{G}\|^a)] \leq A, \forall \mathbf{B} \in \mathbb{R}^{m \times n}.$$

Define the event  $E_t$  to be  $\|\mathbb{E}_{\mathbb{P}_{t, N_t}} \nabla f(\mathbf{B}_t, \xi) - \mathbb{E}_{\mathbb{P}_{\mathbf{B}_t}} \nabla f(\mathbf{B}_t, \xi)\| \leq C_t\}$ , then  $\{E_t\}_{t=0}^{+\infty}$  is a collection of independent events.

**Theorem 8.** *Take  $T > 0$ ,  $\delta > 0$ , choose  $\delta_t$  such that  $\delta > \sum_{t=0}^T \delta_t$ , then with probability at least  $1 - \delta$ , we have that*

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{B}_t)\|^2 \leq \frac{8L_w(f(\mathbf{B}_0) - f^*)}{T} + \frac{8}{T} \sum_{t=0}^{T-1} C_t^2, \quad (7)$$

and thus,

$$\min_{0 \leq t < T} \|\nabla f(\mathbf{B}_t)\|^2 \leq \frac{8L_w(f(\mathbf{B}_0) - f^*)}{T} + \frac{8}{T} \sum_{t=0}^{T-1} C_t^2.$$

**Remark 1.** *Any unitary invariant [4] cross norm [1]  $\|\cdot\|$  shares the same  $C_t$ , and when  $T$  is sufficiently large,  $\sum_{t=0}^{T-1} C_t^2$  becomes the dominant term of the right-hand side. Thus, the right-hand side is approximately invariant among different norms. From proposition 3.12 of [1], the nuclear norm is the largest cross norm over  $\|\cdot\|_2$  and  $\|\cdot\|_2$ . Therefore, when we use a nuclear norm guided steepest descent, (7) is the strongest and the descent is better than when other norms are used because  $\nabla f(\mathbf{B}_t)$  are usually of high rank and  $\|\nabla f(\mathbf{B}_t)\|_{\text{nuc}} > \|\nabla f(\mathbf{B}_t)\|$  for most other unitary invariant cross norms.*



## References

- [1] R. Cochrane. *Tensor Ranks and Norms*. PhD thesis, University of Michigan, 2022. Master’s thesis; accessible via University of Michigan’s Deep Blue repository.
- [2] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.
- [3] L. V. Kantorovich and G. S. Rubinshtein. On a space of totally additive functions. *Vestnik Leningradskogo Universiteta*, 13:52–59, 1958.
- [4] S. Lewis, A. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1/2):173–183, 1995. Received 12 July 1994; revised manuscript December 1995.
- [5] Alexander Shapiro. *On Duality Theory of Conic Linear Problems*, volume 57 of *Nonconvex Optimization and Its Applications*, page 135–165. Springer US, Boston, MA, 2001.

## A Rigorous Proofs of the Theorems

We give proofs only for Theorem 4, Theorem 5, Theorem 6, and Theorem 7, since Theorem 2 and Theorem 3 are special cases of Theorem 4 and Theorem 5, respectively.

### A.1 Proof of Theorem 4

For convenience, we write  $N := N_t, \mathbb{P}_N := \mathbb{P}_{t, N_t}$ . Take  $\mu = \mathbb{Q} - \mathbb{P}_N$ , then  $\mu \in \mathcal{M}(\mathbb{R}^{m \times n})$  and  $\mu(\mathbb{R}^{m \times n}) = 0$ .

$$\begin{aligned} & \sup_{\substack{\mathbb{Q} \in \mathcal{M}_+(\mathbb{R}^{m \times n}) \\ \mathbb{Q}(\mathbb{R}^{m \times n})=1}} \left\{ \mathbb{E}_{\mathbf{G} \sim \mathbb{Q}}[\langle \mathbf{d}, \mathbf{G} \rangle] - \frac{1}{\kappa p} W_1^p(\mathbb{P}_N, \mathbb{Q}) \right\} \\ &= \mathbb{E}_{\mathbf{G} \sim \mathbb{P}_N}[\langle \mathbf{d}, \mathbf{G} \rangle] + \sup_{\substack{\mathbb{Q} \in \mathcal{M}_+(\mathbb{R}^{m \times n}) \\ \mathbb{Q}(\mathbb{R}^{m \times n})=1}} \left\{ \mathbb{E}_{\mathbf{G} \sim \mu}[\langle \mathbf{d}, \mathbf{G} \rangle] - \frac{1}{\kappa p} W_1^p(\mathbb{P}_N, \mathbb{Q}) \right\}. \end{aligned}$$

Consider the normed vector space  $Lip(\mathbb{R}^{m \times n})$  with norm  $\|\cdot\|_{Lip}$  whose value is the smallest Lipschitz constant of that function. Consider a subset of its dual space  $\mathcal{M}(\mathbb{R}^{m \times n})$  that denotes all signed measures on  $\mathbb{R}^{m \times n}$ , and the dual norm is  $\|\cdot\|_{KR}$ . Using Hahn Banach theorem, we know that the dual norm of  $\|\cdot\|_{KR}$  is also  $\|\cdot\|_{Lip}$ .

From Theorem 1,

$$\begin{aligned}
W_1(\mathbb{P}_N, \mathbb{Q}) &= \sup_{\|f\|_{Lip}=1} \left( \int_{\mathbb{R}^{m \times n}} f(\xi) \mathbb{P}_N(d\xi) - \int_{\mathbb{R}^{m \times n}} f(\xi') \mathbb{Q}(d\xi') \right) \\
&= \sup_{\|f\|_{Lip}=1} \int_{\mathbb{R}^{m \times n}} f(\xi) \mu(d\xi) = \|\mu\|_{KR}.
\end{aligned}$$

Thus,

$$\begin{aligned}
& \sup_{\substack{\mathbb{Q} \in \mathcal{M}_+(\mathbb{R}^{m \times n}) \\ \mathbb{Q}(\mathbb{R}^{m \times n})=1}} \left\{ \mathbb{E}_{G \sim \mathbb{Q}}[\langle \mathbf{d}, G \rangle] - \frac{1}{\kappa p} W_1^p(\mathbb{P}_N, \mathbb{Q}) \right\} \\
&= \mathbb{E}_{\mathbf{G} \sim \mathbb{P}_N}[\langle \mathbf{d}, \mathbf{G} \rangle] + \sup_{\substack{\mathbb{Q} \in \mathcal{M}_+(\mathbb{R}^{m \times n}) \\ \mathbb{Q}(\mathbb{R}^{m \times n})=1}} \left\{ \mathbb{E}_{G \sim \mu}[\langle \mathbf{d}, G \rangle] - \frac{1}{\kappa p} \|\mu\|_{KR}^p \right\}.
\end{aligned}$$

While

$$\sup_{\substack{\mathbb{Q} \in \mathcal{M}_+(\mathbb{R}^{m \times n}) \\ \mathbb{Q}(\mathbb{R}^{m \times n})=1}} \left\{ \mathbb{E}_{G \sim \mu}[\langle \mathbf{d}, G \rangle] - \frac{1}{\kappa p} \|\mu\|_{KR}^p \right\}$$

is just the conjugate function of  $\frac{1}{\kappa p} \|\cdot\|_{KR}^p$  which equals to  $\frac{\kappa}{q} \|\cdot\|_{Lip}^q$ , where  $\frac{1}{p} + \frac{1}{q} = 1$ . We also know that  $\frac{\kappa}{q} \|\langle \mathbf{d}, \cdot \rangle\|_{Lip}^q = \frac{\kappa}{q} \|\mathbf{d}\|_{op}^q$ . Therefore, we end up with the conclusion in the theorem.

$$\begin{aligned}
& \sup_{\substack{\mathbb{Q} \in \mathcal{M}_+(\mathbb{R}^{m \times n}) \\ \mathbb{Q}(\mathbb{R}^{m \times n})=1}} \left\{ \mathbb{E}_{G \sim \mathbb{Q}}[\langle \mathbf{d}, G \rangle] - \frac{1}{\kappa p} W_1^p(\mathbb{P}_N, \mathbb{Q}) \right\} \\
&= \mathbb{E}_{G \sim \mathbb{P}_N}[\langle \mathbf{d}, G \rangle] + \frac{\kappa}{q} \|\mathbf{d}\|_{op}^q.
\end{aligned}$$

□

## B Proof of Theorem 5

Assume that  $\mathbf{d} = c\mathbf{M}$ , where  $\|\mathbf{M}\|_{op} = 1$ , then our goal is:

$$\min_{c \geq 0, \mathbf{M} \in \mathbb{R}^{m \times n}} \left( c \langle \mathbf{M}, \overline{G}_t \rangle + \frac{1}{q} \kappa c^q \right).$$

By Neumann's Inequality,

$$\min_{\mathbf{d} \in \mathbb{R}^{m \times n}} \langle \mathbf{M}, \overline{G}_t \rangle = -\|\overline{G}_t\|_{nuc}.$$

The minimal value is achieved if and only if  $\mathbf{M} = -U_t V_t^T$ . When the minimal value is achieved, it suffices to minimize

$$-c\|\overline{G}_t\|_{nuc} + \frac{1}{q}\kappa c^q,$$

and the optimal  $c$  is  $(\frac{\|\overline{G}_t\|_{nuc}}{\kappa})^{1/(q-1)}$  naturally.  $\square$

## C Proof of Theorem 6

For simplicity, write  $N = N_t$ ,  $\mathbb{P}_N = \mathbb{P}_{t, N_t}$ . We follow the notations in [5], take  $X = (\mathcal{M}(\mathbb{R}^{m \times n}))^N$  to be the space of  $N$ -tuples of signed finite measures on  $\mathbb{R}^{m \times n}$ . Take  $C$  as a convex conic subset of  $X$  which contains all nonnegative measures. Take  $Y = \mathbb{R}^{N+1}$  and  $K = \mathbb{R}_{\leq 0} \times \{0\}^N$ .

Therefore,  $X^* = \mathcal{L}^\infty(\mathbb{R}^{m \times n})$ ,  $C^* = \{f \in C : f \geq 0\}$ ,  $Y^* = \mathbb{R}^{N+1}$ ,  $K^* = \mathbb{R}_{\leq 0} \times \mathbb{R}^N$ . We also take  $b = (-(C + L\|\mathbf{d}\|_*), -1, \dots, -1)^T \in \mathbb{R}^{N+1}$ .

There exists a nonnegative measure  $\pi$  on  $\mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n}$  whose marginal distribution for the first  $m$  components is  $\mathbb{P}_N$  and the marginal distribution for the second  $m$  components is  $\mathbb{Q}$ , such that

$$\mathbb{E}_{(\xi, \xi') \sim \pi} \|\xi - \xi'\| = W(\mathbb{P}_N, \mathbb{Q}).$$

We denote

$$\mathbf{q} = (\mathbb{Q}^1, \mathbb{Q}^2, \dots, \mathbb{Q}^N)^T \in X.$$

The first stage problem of (P)

$$\sup_{W(\mathbb{P}_N, \mathbb{Q}) \leq C + L\|\mathbf{d}\|_*} \mathbb{E}_{\mathbb{Q}}[\langle \mathbf{d}, \nabla_{\mathbf{B}} f_{\mathbf{x}}(\mathbf{B}) \rangle]$$

is equivalent to

$$\begin{cases} \sup_{\Pi \in M(\mathbb{R}^{m \times n}, \mathbb{R}^{m \times n})} \int_{\mathbb{R}^{m \times n}} \langle \mathbf{d}, \mathbf{G} \rangle \Pi(dG, \mathbb{R}^{m \times n}), \\ \text{s.t.} & \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}^{m \times n}} \|\mathbf{G} - \mathbf{G}_i\| \mathbb{Q}^i(dG) \leq C + L\|\mathbf{d}\|_{op}, \\ & \int_{\mathbb{R}^{m \times n}} \mathbb{Q}^i(dG) = 1, \end{cases} \quad (8)$$

We write  $\mathbf{G} = \nabla_{\mathbf{B}} f_{\mathbf{x}}(\mathbf{B})$ ,  $\mathbf{G}_i = \nabla_{\mathbf{B}} f_{\mathbf{x}_i}(\mathbf{B})$ ,  $c = (\frac{1}{N} \langle \mathbf{d}, \mathbf{G} \rangle)_{i=1}^N$ ,  $A\mathbf{q} = (A^0\mathbf{q}, A^1\mathbf{q}, \dots, A^N\mathbf{q})^T$ , where  $A^i\mathbf{q} = \int_{\mathbb{R}^{m \times n}} \mathbb{Q}^i(dG)$ ,  $A^0\mathbf{q} = \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}^{m \times n}} \|G - G_i\| \mathbb{Q}^i(dG)$ .

The (8) is equivalent to

$$\begin{cases} \sup_{\mathbb{Q} \in C} \langle c, \mathbf{q} \rangle, \\ \text{s.t.} & A\mathbf{q} + b \in K. \end{cases} \quad (9)$$

We now turn to look at its dual problem:

$$\begin{cases} \sup_{y^* \in K^*} \langle b, y^* \rangle, \\ \text{s.t.} & A^*y^* + c \in C^*. \end{cases} \quad (10)$$

Assume that  $y^* = (\lambda, s_1, s_2, \dots, s_N)$ , then

$$\begin{aligned}
\langle y^*, A\mathbf{q} \rangle &= \frac{\lambda}{N} \sum_{i=1}^N \int_{\mathbb{R}^{m \times n}} \|\mathbf{G} - \mathbf{G}_i\| \mathbb{Q}^i(dG) + \sum_{i=1}^N s^i \int_{\mathbb{R}^{m \times n}} \mathbb{Q}^i(dG) \\
&= \int_{\mathbb{R}^{m \times n}} \sum_{i=1}^N \left( \frac{\lambda}{N} \|\mathbf{G} - \mathbf{G}_i\| + s^i \right) \mathbb{Q}^i(dG) \\
&= \langle A^* y^*, x \rangle.
\end{aligned}$$

Hence,  $A^* y^* = (\frac{\lambda}{N} \|\mathbf{G} - \mathbf{G}_i\| + s^i)_{i=1}^N$ . Since  $A^* y^* + c \in C^*$ ,

$$0 \geq A^* y^* + c = \left( \frac{\lambda}{N} \|\mathbf{G} - \mathbf{G}_i\| + s^i + \frac{1}{N} \langle \mathbf{d}, \mathbf{G}_i \rangle \right), \forall i. \quad (11)$$

(11) is equivalent to that

$$\begin{aligned}
-s^i &\geq \frac{1}{N} \sup_{\mathbf{G}} (\langle \mathbf{d}, \mathbf{G} \rangle + \lambda \|\mathbf{G} - \mathbf{G}_i\|) \\
&= \frac{1}{N} \sup_{\mathbf{G}} (\langle \mathbf{d}, \mathbf{G} - \mathbf{G}_i \rangle - |\lambda| \|\mathbf{G} - \mathbf{G}_i\|) + \frac{\langle \mathbf{d}, \mathbf{G}_i \rangle}{N} \\
&= \frac{1}{N} (|\lambda| \|\cdot\|)^*(\mathbf{d}) + \frac{\langle \mathbf{d}, \mathbf{G}_i \rangle}{N} \\
&= \frac{\langle \mathbf{d}, \mathbf{G}_i \rangle}{N} + \begin{cases} 0, & \|\mathbf{d}\|_* \leq -\lambda, \\ +\infty, & \text{otherwise.} \end{cases}
\end{aligned}$$

Notice that  $\langle c, \mathbf{q} \rangle = -\lambda(C + L\|\mathbf{d}\|) - \sum_{i=1}^N s^i$ , to minimize this, take  $-\lambda = \|\mathbf{d}\|_*$  and  $-s^i = \frac{\langle \mathbf{d}, \mathbf{G}_i \rangle}{N}$ .

(10) is now written as

$$\min_{\mathbf{d}} \frac{1}{N} \sum_{i=1}^N \langle \mathbf{d}, \mathbf{G}_i \rangle + (C + L\|\mathbf{d}\|_*) \|\mathbf{d}\|_*.$$

To show strong duality in Proposition 3.4 of [4], notice that  $A(C) = \mathbb{R}_{\geq 0}^{N+1}$ ,  $A(C) - K = \mathbb{R}_{\geq 0}^{N+1}$ ,  $-b \in \text{int}(A(C) - K)$ .

□

## D Proof of Theorem 7

The methodology is very similar to the proof of Theorem 5. We repeat the proof here for completeness.

Assume that  $\mathbf{d} = c\mathbf{M}$ , where  $\|\mathbf{M}\|_{op} = 1$ , then our goal is:

$$\min_{c \geq 0, \mathbf{M} \in \mathbb{R}^{m \times n}} (c \langle \mathbf{M}, \overline{\mathbf{G}}_i \rangle + Cc + Lc^2).$$

By Neumann's Inequality,

$$\min_{\mathbf{d} \in \mathbb{R}^{m \times n}} \langle \mathbf{M}, \overline{\mathbf{G}}_t \rangle = -\|\overline{\mathbf{G}}_t\|_{nuc}.$$

The minimal value is achieved if and only if  $\mathbf{M} = -U_t V_t^T$ . When the minimal value is achieved, it suffices to minimize

$$(C - \|\mathbf{G}_t\|_{nuc})c + Lc^2, c \geq 0.$$

and the minimal value is  $\frac{\max\{0, \|\mathbf{G}_t\|_{nuc} - C\}}{2L}$  naturally.  $\square$

## E Proof of Lemma 1

Take any  $\xi \in \Xi$ ,  $\mathbf{B} \in \mathbb{R}^{m \times n}$ , then

$$\nabla f(\mathbf{B}, \xi) = \frac{\partial f(\mathbf{B}, \xi)}{\partial \mathbf{B}} = \frac{\partial f(\mathbf{B}, \xi)}{\partial \mathbf{a}^l} \mathbf{h}^{l-1^T},$$

is of rank one.  $\square$

## F Proof of Theorem 8

**Lemma 1.** *If event  $E_t$  happens, then  $\forall \mathbf{d} \in \mathbb{R}^{m \times n}$ ,*

$$f(\mathbf{B}_t + \mathbf{d}) - f(\mathbf{B}_t) \leq \sup_{W(\mathbb{P}_{t, N_t}, \mathbb{Q}) \leq C_t + L_w \|\mathbf{d}\|_*} \mathbb{E}_{\mathbb{Q}} \langle \mathbf{d}, \mathbf{G} \rangle.$$

*Proof.* By the mean-value problem, there exists  $\theta \in (0, 1)$ , such that

$$f(\mathbf{B}_t + \mathbf{d}) - f(\mathbf{B}_t) = \langle \nabla f(\mathbf{B}_t + \theta \mathbf{d}), \mathbf{d} \rangle = \mathbb{E}_{\mathbf{G} \sim \mathbb{P}_{\mathbf{B}_t + \theta \mathbf{d}}} \langle \mathbf{G}, \mathbf{d} \rangle.$$

Since

$$W(\mathbb{P}_{\mathbf{B}_t + \theta \mathbf{d}}, \mathbb{P}_{t, N_t}) \leq W(\mathbb{P}_{\mathbf{B}_t + \theta \mathbf{d}}, \mathbb{P}_{\mathbf{B}_t}) + W(\mathbb{P}_{t, N_t}, \mathbb{P}_{\mathbf{B}_t}) \leq L_w \|\mathbf{d}\|_* + C_t,$$

we have that,

$$f(\mathbf{B}_t + \mathbf{d}) - f(\mathbf{B}_t) = \mathbb{E}_{\mathbf{G} \sim \mathbb{P}_{\mathbf{B}_t + \theta \mathbf{d}}} \langle \mathbf{G}, \mathbf{d} \rangle \leq \sup_{W(\mathbb{P}_{t, N_t}, \mathbb{Q}) \leq C_t + L_w \|\mathbf{d}\|_*} \mathbb{E}_{\mathbb{Q}} \langle \mathbf{d}, \mathbf{G} \rangle.$$

$\square$

**Lemma 2.** *On the event  $E_t$ , if  $\mathbf{B}_{t+1} = \mathbf{B}_t + \mathbf{d}_t$ , then*

$$f(\mathbf{B}_{t+1}) - f(\mathbf{B}_t) \leq -\frac{(\|\overline{\mathbf{G}}_t\| - C_t)_+^2}{4L_w}.$$

*Proof.* Since

$$f(\mathbf{B}_t + \mathbf{d}_t) - f(\mathbf{B}_t) \leq \sup_{W(\mathbb{P}_{t, N_t}, \mathbb{Q}) \leq C_t + L_w \|\mathbf{d}\|_*} \mathbb{E}_{\mathbb{Q}} \langle \mathbf{d}_t, \mathbf{G} \rangle = \langle \mathbf{d}_t, \overline{\mathbf{G}}_t \rangle + C_t \|\mathbf{d}_t\|_* + L_w \|\mathbf{d}_t\|_*^2.$$

Since  $\mathbf{d}_t$  is taken to be the infimum of the right-hand side,

$$f(\mathbf{B}_t + \mathbf{d}_t) - f(\mathbf{B}_t) \leq \inf_{c \geq 0} c(-\|\overline{\mathbf{G}}_t\| + C_t) + L_w c^2 = -\frac{(\|\overline{\mathbf{G}}_t\| - C_t)_+^2}{4L_w}.$$

□

**Lemma 3.**

$$\|\overline{\mathbf{G}}_t - \nabla f(\mathbf{B}_t)\| \leq C_t.$$

*Proof.* Take arbitrarily  $\mathbf{d} \in \mathbb{R}^{m \times n}$ ,  $\|\mathbf{d}\|_* = 1$ , we have by Theorem 1 that

$$\begin{aligned} \langle \mathbf{d}, \overline{\mathbf{G}}_t - \nabla f(\mathbf{B}_t) \rangle &= \int_{\mathbb{R}^{m \times n}} \langle \mathbf{d}, \mathbf{G} \rangle \mathbb{P}_{t, N_t}(\mathbf{dG}) - \int_{\mathbb{R}^{m \times n}} \langle \mathbf{d}, \mathbf{G} \rangle \mathbb{P}_{\mathbf{B}_t}(\mathbf{dG}) \\ &\leq W(\mathbb{P}_{t, N_t}, \mathbb{P}_{\mathbf{B}_t}) \\ &\leq C_t. \end{aligned}$$

Therefore,  $\|\overline{\mathbf{G}}_t - \nabla f(\mathbf{B}_t)\| \leq C_t$ .

□

**Lemma 4.** If  $x, y \geq 0$ , then  $(x - y)_+^2 \geq \frac{1}{2}x^2 - y^2$ .

*Proof.* Case 1:  $x \geq y$ ,

In this case,  $(x - y)_+ = x - y$ , so the left side is  $(x - y)^2 = x^2 - 2xy + y^2$ . The inequality becomes  $x^2 - 2xy + y^2 \geq \frac{1}{2}x^2 - y^2$ . Rearranging terms gives  $\frac{1}{2}x^2 - 2xy + 2y^2 \geq 0$ . Multiplying through by 2 yields  $x^2 - 4xy + 4y^2 \geq 0$ , or  $(x - 2y)^2 \geq 0$ . This is always true, with equality when  $x = 2y$ .

Case 2:  $x < y$ ,

In this case,  $(x - y)_+ = 0$ , so the left side is 0. The inequality becomes  $0 \geq \frac{1}{2}x^2 - y^2$ , or  $y^2 \geq \frac{1}{2}x^2$ . Since  $x < y$  and both are non-negative,  $y^2 > x^2 \geq \frac{1}{2}x^2$ , so the inequality holds strictly.

□

**Lemma 5.**

$$f(\mathbf{B}_t + \mathbf{d}_t) - f(\mathbf{B}_t) \leq -\frac{1}{8L_w} \|\nabla f(\mathbf{B}_t)\|^2 + \frac{1}{L_w} C_t^2. \quad (12)$$

*Proof.*

$$\begin{aligned} f(\mathbf{B}_t + \mathbf{d}_t) - f(\mathbf{B}_t) &\leq -\frac{(\|\overline{\mathbf{G}}_t\| - C_t)_+^2}{4L_w} \leq -\frac{(\|\nabla f(\mathbf{B}_t)\| - 2C_t)_+^2}{4L_w} \\ &\leq -\frac{\frac{1}{2}\|\nabla f(\mathbf{B}_t)\|^2 - 4C_t^2}{4L_w} \\ &= -\frac{1}{8L_w} \|\nabla f(\mathbf{B}_t)\|^2 + \frac{1}{L_w} C_t^2. \end{aligned}$$

□

Now Theorem 8 follows by taking the summation of (12) from 0 to  $T - 1$ , dividing both sides by  $T$ , and rearranging.