

608 A Additonnal Experiments

609 A.1 Mini-batch DRAG.

610 As for Vanilla SGD, we can take advantage of GPU parallelization and replace the gradient estimator
 611 using one sample $X \sim \mu$

$$\nabla_{\mathbf{g}} h_{\varepsilon}(X, \mathbf{g})$$

612 by a mini-batch estimator, using $n_b \geq 1$ i.i.d samples X_1, \dots, X_{n_b} samples of the source measure at
 613 once

$$\frac{1}{n_b} \sum_{k=0}^{n_b} \nabla_{\mathbf{g}} h_{\varepsilon}(X_k, \mathbf{g}). \quad (7)$$

614 Of course, no matter the choice n_b , (7) defines an unbiased estimator of $\nabla H_{\varepsilon}(\mathbf{g})$.

615 Using a mini-batch of size n_b , we suggest multiplying γ_1 by $\sqrt{n_b}$, as is usual with mini-batch SGD.
 616 The following figure shows the acceleration due to mini-batching on Example 1, 2 and 3, while
 617 maintaining the same computational time when using a GPU. Indeed, each mini-batch estimator has
 618 an error an order of magnitude lower than the non-batched ones, even with a small mini-batch size of
 619 $n_b = 16$.

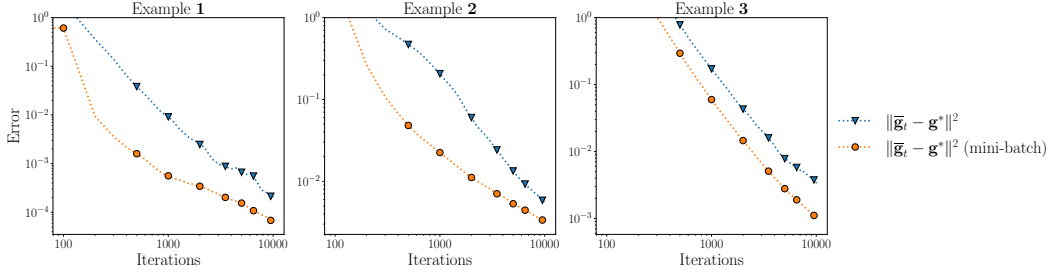


Figure 6: Comparison of the non mini-batched and mini-batched estimators on Example 1, 2 and 3, $n_b = 16$.

620 A.2 Weighted Averaging: Maintaining a better trade-off between averaged and non-averaged 621 iterations.

622 It is well known that the averaged algorithm can suffer from bad initialization. One strategy to over-
 623 come this is weighted averaging [39]. Namely, we replace the averaged estimator $\bar{\mathbf{g}}_t = \frac{1}{t+1} \sum_{k=0}^t \mathbf{g}_k$,
 624 by

$$\bar{\mathbf{g}}_t^{(\omega)} := \frac{1}{\sum_{k=0}^t \log(k+1)^\omega} \sum_{k=0}^t \log(k+1)^\omega \mathbf{g}_k,$$

625 with a parameter $\omega > 0$. The parameter ω balances the weights assigned to the estimators \mathbf{g}_k . As
 626 ω increases, greater importance is given to the more recent estimates, while we retrieve $\bar{\mathbf{g}}_t$ when ω
 627 goes to 0. As for the usual averaged estimators, we can perform the weighted average online, without
 628 having to store all the iterates, with the recursion

$$\bar{\mathbf{g}}_{t+1}^{(\omega)} = \left(1 - \frac{\ln(t+1)^\omega}{\sum_{k=0}^t \ln(k+1)^\omega} \right) \bar{\mathbf{g}}_t^{(\omega)} + \frac{\ln(t+1)^\omega}{\sum_{k=0}^t \ln(k+1)^\omega} \mathbf{g}_{t+1}.$$

629 It is important to note that $\bar{\mathbf{g}}_t^{(\omega)}$ will have the same asymptotic convergence guarantees as $\bar{\mathbf{g}}_t$.

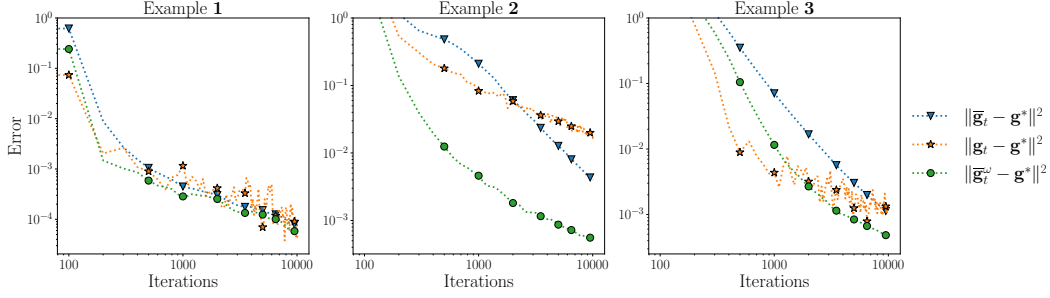


Figure 7: Comparison between \mathbf{g}_t , $\bar{\mathbf{g}}_t$ and $\bar{\mathbf{g}}_t^{(\omega)}$ on Examples 1, 2 and 3, with $\omega = 2$

As illustrated in Figure 7, the weighted average estimator consistently outperforms $\bar{\mathbf{g}}_t$, achieving orders of magnitude better performance in Examples 2 and 3. Note that a mini-batch size of 16 was used for all experiments, each repeated 10 times.

A.3 DRAG compared to Adam.

We compare here the performance of our algorithm DRAG to that of the Adam algorithm [30], on Example 1, with $M \in 200, 2000$. The experiment was repeated 10 times. For this comparison, we fixed the parameters of DRAG to $(\sqrt{M}, 1/3, 2/3)$ and ran the algorithm for $t = 10^5$ iterations. The parameters for Adam were set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\lambda = 10^{-3}$ (learning rate/weight decay). As shown in Figure 8, DRAG clearly outperforms Adam on this example, particularly in the early iterations and as the number of points increases.

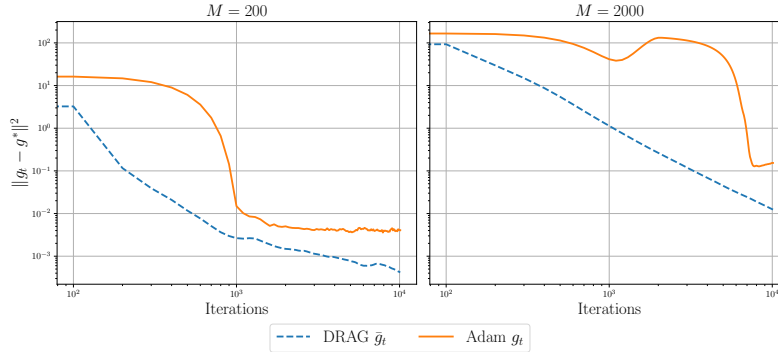


Figure 8: Comparison of DRAG with Adam on Example 1, for different values of M .

A.4 DRAG compared to non regularized ASGD

As discussed in the numerical section, we observed empirically that all methods, including the non-regularized projected ASGD, eventually converge to the true solution given a sufficient number of iterations. In Figure 9, we report additional experiments with a larger iteration budget to confirm this behavior. Notably, even the non-regularized ASGD converges, albeit much more slowly. In contrast, DRAG converges significantly faster and achieves higher accuracy earlier in the optimization process. Among the DRAG variants, we observe that choices of $a \in 0.2, 0.4$ yield the best performance, which aligns with our theoretical analysis suggesting that $a = \frac{1}{3}^-$ achieves the optimal convergence rate in this setting, since $b = \frac{2}{3}$. These results reinforce our claim from the main text that, while all schemes converge given enough iterations, DRAG remains consistently one to two orders of magnitude more accurate due to its improved early-stage performance (see Appendix A).

Note also that the convergence of the non-regularized ASGD is encouraging, as it highlights that, with a sufficiently large number of steps, the effect of vanishing regularization is not a deterrent. Indeed, as $t \rightarrow \infty$, the regularized gradient becomes numerically indistinguishable from the non-regularized one, essentially corresponding to the difference between a tempered softmax with temperature ε and

an argmax. This further supports the view that DRAG serves as an effective acceleration mechanism in the early stages of optimization and that by decreasing the regularization, we will not hit a plateau.

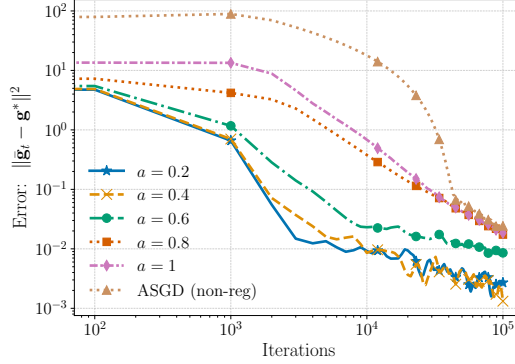


Figure 9: Comparison of DRAG with different a and non-regularized ASGD

B Proofs

Additional notations.

For any $c > 0$ we define the function $t \mapsto \Psi_c(t)$ such that

$$\sum_{t=1}^T t^{-c} \leq \Psi_c(T) := \begin{cases} 1 + \ln(T+1) & \text{if } c = 1, \\ \frac{2c-1}{c-1} & \text{if } c > 1, \\ 1 + \frac{1}{1-c}(T+1)^{1-c} & \text{if } c < 1. \end{cases} \quad (8)$$

For a sequence $(u_t)_{t \in \mathbb{N}}$, if $\frac{t}{2} \notin \mathbb{N}$, $u_{\frac{t}{2}}$ must be understood as $u_{\lceil \frac{t}{2} \rceil}$.

In all the sequel, we note

$$\Delta_t = \|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\|^2.$$

Remark that the dependence in t is both in the estimator \mathbf{g}_t and the optimizer $\mathbf{g}_{\varepsilon_t}^*$. We also recall that we note $D_C := \sup_{\mathbf{g}, \mathbf{g}' \in C} \|\mathbf{g} - \mathbf{g}'\| < \infty$.

B.1 Proof of Theorem 1: Convergence rate of the non averaged iterates.

Proof. Using Lemma 3, for any $t \geq t_{a,\alpha}$, we have

$$\Delta_{t+1} \leq \Delta_t - 2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle + 5\gamma_{t+1}^2.$$

Let \mathcal{F}_t denote the filtration generated by the samples $X_1, \dots, X_t \stackrel{\text{iid}}{\sim} \mu$, that is $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ and taking the conditional expectation, we have

$$\mathbb{E}[\Delta_{t+1} | \mathcal{F}_t] \leq \Delta_t - 2\gamma_{t+1} \langle \nabla H_{\varepsilon_t}(\mathbf{g}_t), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle + 5\gamma_t^2. \quad (9)$$

Using Lemma 2 on the restricted strong convexity of H_{ε_t} , we have

$$\langle \nabla H_{\varepsilon_t}(\mathbf{g}_t), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle \geq \rho_*(1 - e^{-1}) \|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\|^2 \mathbb{1}_{\|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\| \leq \varepsilon_t/2}.$$

Therefore, we have

$$\mathbb{E}[\Delta_{t+1} | \mathcal{F}_t] \leq [1 - 2\rho_*(1 - e^{-1})\gamma_{t+1}] \Delta_t + \left[2\rho_*(1 - e^{-1}) \mathbb{1}_{\|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\| \geq \varepsilon_t/2} \right] \gamma_{t+1} \Delta_t + 5\gamma_{t+1}^2. \quad (10)$$

Using Proposition 2, for all p and $\beta \in (0, 1)$, there exists $C_{\beta,p}$ such that for all $t \geq 0$, $\mathbb{E}[\Delta_t^p] \leq C_{\beta,p} \frac{\gamma_t^p \varepsilon_t^{\beta p}}{\varepsilon_t^p} \leq C_{\beta,p} t^{-bp+a(1-\beta)p}$. Therefore,

$$\begin{aligned} \mathbb{E}[\mathbb{1}_{\|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\| \geq \varepsilon_t/2}] &= \mathbb{E}[\mathbb{1}_{\|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\|^{2p} \geq \varepsilon_t^{2p}/2^p}] \\ &\leq C_{\beta,p} t^{-bp+ap(3-\beta)} \quad \text{by Markov's inequality} \\ &\leq C_{\beta, \frac{4b}{b-a(3-\beta)}} t^{-4b} \quad \text{taking } p = \frac{4b}{b-a(3-\beta)}, \text{ with } a(3-\beta) < b. \end{aligned} \quad (11)$$

Note that, since $2a < b$, we can always choose β such that the inequality $a(3-\beta) < b$ holds. Using the fact that $\Delta_t \leq D_C^2$ and taking the expectation in (10), we obtain

$$\mathbb{E}[\Delta_{t+1}] \leq [1 - 2\rho_*(1 - e^{-1})\gamma_{t+1}] \mathbb{E}[\Delta_t] + 5\gamma_{t+1}^2 + C_{\beta, \frac{4b}{b-a(3-\beta)}} t^{-5b} D_C^2.$$

Let $t_\gamma := \min\{t, 2\lambda\gamma_{t+1} \leq 1\}$ and $t_0 := \max\{t_{a,\alpha}, t_\gamma\}$, we apply Proposition 5 to obtain

$$\mathbb{E}[\Delta_t] \leq \exp\left(-2\lambda \sum_{i=t_0+1}^t \gamma_i\right) \left(D_C^2 + \sum_{k=t_0}^t 5\gamma_k^2\right) + \frac{5}{2\lambda} \gamma_{\frac{t}{2}-1} + o(\gamma_t). \quad (12)$$

Applying Corollary 2, the exponential product converges exponentially fast to 0 and an asymptotic comparison gives

$$\mathbb{E}[\Delta_t] \leq \frac{5}{2\rho^*(1 - e^{-1})} \gamma_{\frac{t}{2}-1} + o(\gamma_t) \lesssim \frac{\gamma_t}{\rho^*}.$$

We conclude by using Proposition 1, using the bound $\|\mathbf{g}_{\varepsilon_t}^* - \mathbf{g}^*\| \lesssim \varepsilon_t^{1+\alpha'}$.

□

Proposition 3. Under the same assumptions as in Theorem 1, we have for any $\alpha' \in (0, \alpha)$

$$\mathbb{E}[\|\mathbf{g}_t - \mathbf{g}^*\|^2] \lesssim \frac{1}{\rho_*^2 \cdot t^{2b}} + \frac{1}{t^{4a+4a\alpha'}}, \quad t \geq 1.$$

Remark: Note that this proposition directly proves Theorem 1, but we decided to split them, to have a cleaner proof of Theorem 1.

Proof. We begin by squaring equation (13) of Lemma 3. For $t \geq t_{a,\alpha}$, where $t_{a,\alpha}$ is defined in (15), we have

$$\begin{aligned} \Delta_{t+1}^2 &\leq (\Delta_t - 2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle + 5\gamma_{t+1}^2)^2 \\ &\leq \Delta_t^2 + 4\gamma_{t+1}^2 \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle^2 + 25\gamma_{t+1}^4 \\ &\quad - \underbrace{4\Delta_t \gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle}_{=:A} + 10\Delta_t \gamma_{t+1}^2 - \underbrace{20\gamma_{t+1}^3 \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle}_{=:B}. \end{aligned}$$

Taking the conditional and using Lemma 2, we have

$$\mathbb{E}[A \mid \mathcal{F}_t] \geq 4\Delta_t^2 \rho_*(1 - e^{-1})\gamma_{t+1} \mathbb{1}_{\|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}\| \leq \varepsilon_t/2}.$$

We also use the simple bound

$$\mathbb{E}[B \mid \mathcal{F}_t] \geq 0.$$

These two inequalities lead to

$$\begin{aligned} \mathbb{E}[\Delta_t^2 \mid \mathcal{F}_t] &\leq [1 - 4\rho_*(1 - e^{-1})\gamma_{t+1}] \Delta_t^2 + 4\rho_*(1 - e^{-1})\gamma_{t+1} \mathbb{1}_{\|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}\| \leq \varepsilon_t/2} \\ &\quad + 4\gamma_{t+1}^2 \mathbb{E}[\langle \nabla h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle^2 \mid \mathcal{F}_t] + 25\gamma_{t+1}^4 + 5\Delta_t \gamma_{t+1}^2. \end{aligned}$$

687 Using that the gradient norm is bounded by two, we apply the Cauchy-Schwarz inequality to obtain

$$4\gamma_{t+1}^2 \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle^2 \leq 16\gamma_{t+1}^2 \|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\|^2 \leq 16\Delta_t \gamma_{t+1}^2.$$

688 Applying Hölder's inequality yields

$$\begin{aligned} 21\Delta_t \gamma_{t+1}^2 &\leq \left(\Delta_t \sqrt{2\rho_*(1-e^{-1})} \frac{1}{\sqrt{2\rho_*(1-e^{-1})}} 21\gamma_{t+1} \right) \gamma_{t+1} \\ &\leq \gamma_{t+1} \Delta_t^2 \rho_*(1-e^{-1}) + \frac{21^2}{4\rho_*(1-e^{-1})} \gamma_{t+1}^3. \end{aligned}$$

689 Summing up these inequalities, we obtain

$$\begin{aligned} \mathbb{E}[\Delta_{t+1}^2 \mid \mathcal{F}_t] &\leq [1 - 3\rho_*(1-e^{-1})\gamma_{t+1}] \Delta_t^2 + 4\rho_*(1-e^{-1})\gamma_{t+1} \mathbb{1}_{\|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\| \geq \varepsilon_t/2} \\ &\quad + \frac{21^2}{4\rho_*(1-e^{-1})} \gamma_{t+1}^3 + 25\gamma_{t+1}^4. \end{aligned}$$

690 Similarly to the case $p = 1$, using that $\mathbb{P}[\|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\| \geq \varepsilon_t/2] \leq C_{\beta, \frac{4b}{b-a(3-\beta)}} t^{-4b}$ by (11) and that

691 $\Delta_t^2 \leq D_C^4$ for all t , taking the expectation yields

$$\mathbb{E}[\Delta_{t+1}^2] \leq (1 - 3\lambda\gamma_{t+1})\mathbb{E}[\Delta_t^2] + \frac{21^2}{4\lambda} \gamma_{t+1}^3 + 25\gamma_{t+1}^4 + 4\lambda C_{\beta, \frac{4b}{b-a(3-\beta)}} t^{-5b} D_C^4.$$

692 Again, as in the case $p = 1$, applying Proposition 5 and Corollary 2 and using that $\|\mathbf{g}_{\varepsilon_t}^* - \mathbf{g}^*\| \lesssim \varepsilon_t^{1+\alpha'}$
693 concludes the proof.

694 □

695 **Lemma 3.** *Under the assumptions of Theorem 1, there exists a finite time $t_{a,\alpha}$, depending on a and*
696 *α , such that for all $t \geq t_{a,\alpha}$, we have*

$$\Delta_{t+1} \leq \Delta_t - 2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle + 5\gamma_{t+1}^2. \quad (13)$$

697 *Proof.* By definition of the gradient step at time $t + 1$ and since $\mathbf{g}_{\varepsilon_{t+1}}^* \in \mathcal{C}$, we have

$$\begin{aligned} \Delta_{t+1} &= \|\mathbf{g}_{t+1} - \mathbf{g}_{\varepsilon_{t+1}}^*\|^2 \\ &= \|\text{Proj}_{\mathcal{C}}(\mathbf{g}_t - \gamma_{t+1} \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1})) - \mathbf{g}_{\varepsilon_{t+1}}^*\|^2 \\ &\leq \|\mathbf{g}_t - \gamma_{t+1} \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}) - \mathbf{g}_{\varepsilon_{t+1}}^*\|^2. \end{aligned}$$

698 Then, incorporating the change of optimum between time t and $t + 1$, we get

$$\begin{aligned} \Delta_{t+1} &\leq \|\mathbf{g}_t - \gamma_{t+1} \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}) - \mathbf{g}_{\varepsilon_t}^* + \mathbf{g}_{\varepsilon_t}^* - \mathbf{g}_{\varepsilon_{t+1}}^*\|^2 \\ &\leq \|\mathbf{g}_t - \gamma_{t+1} \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}) - \mathbf{g}_{\varepsilon_t}^*\|^2 + 2 \left\langle \mathbf{g}_t - \gamma_{t+1} \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}) - \mathbf{g}_{\varepsilon_t}^*, \mathbf{g}_{\varepsilon_t}^* - \mathbf{g}_{\varepsilon_{t+1}}^* \right\rangle \\ &\quad + \|\mathbf{g}_{\varepsilon_t}^* - \mathbf{g}_{\varepsilon_{t+1}}^*\|^2. \end{aligned}$$

699 Using Corollary 2.2 in [18] (see Proposition 1), there exists $K_0 > 0$ such that for any $\alpha' \in]0, \alpha[$

$$\|\mathbf{g}_{\varepsilon_t}^* - \mathbf{g}_{\varepsilon_{t+1}}^*\| \leq K_0 \varepsilon_t^{\alpha'} (\varepsilon_t - \varepsilon_{t+1}) \leq K_0 t^{-a\alpha'} (t^{-a} - (t+1)^{-a}) \leq aK_0 t^{-(1+a+a\alpha')}. \quad (14)$$

700 For clarity, we define $r_t := aK_0 t^{-(1+a+a\alpha')}$ and $R_t := (2D_C + 2\gamma_{t+1} + r_t)r_t$.

701 Using that for all t , $\mathbf{g}_t \in \mathcal{C}$, and that for all $x \in \mathbb{R}^d$, $\mathbf{g} \in \mathbb{R}^M$, $\|\nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}, x)\| \leq 2$, we obtain

$$\begin{aligned} \Delta_{t+1} &\leq \|\mathbf{g}_t - \gamma_{t+1} \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}) - \mathbf{g}_{\varepsilon_t}^*\|^2 + (2D_C + 2\gamma_{t+1}) \|\mathbf{g}_{\varepsilon_t}^* - \mathbf{g}_{\varepsilon_{t+1}}^*\| + \|\mathbf{g}_{\varepsilon_t}^* - \mathbf{g}_{\varepsilon_{t+1}}^*\|^2 \\ &\leq \|\mathbf{g}_t - \gamma_{t+1} \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}) - \mathbf{g}_{\varepsilon_t}^*\|^2 + R_t \\ &\leq \|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\|^2 - 2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle + \gamma_{t+1}^2 \|\nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1})\|^2 + R_t \\ &\leq \Delta_t - 2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle + 4\gamma_{t+1}^2 + R_t. \end{aligned}$$

702 Note that, since we have $1 + a + a\alpha > 2b$, we can also take $\alpha' \in]0, \alpha[$ such that $1 + a + a\alpha' > 2b$.
 703 Consequently, the sequence R_t/γ_t^2 is decreasing and tends to 0. For conciseness, we note

$$t_{a,\alpha} := \min \{t \geq 1 : R_t \leq \gamma_t^2\}. \quad (15)$$

704 For any $t \geq t_{a,\alpha}$, we then obtain the following upper bound of Δ_{t+1} in terms of Δ_t and the gradient
 705 direction:

$$\Delta_{t+1} \leq \Delta_t - 2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^* \rangle + 5\gamma_{t+1}^2.$$

706

□

707 B.2 Proof of Theorem 2: Convergence rate of DRAG

708 *Proof.* We start with a decomposition of the gradient step, similar to [26]. By abuse of notation, we
 709 note

$$\nabla_*^2 := \nabla^2 H_0(\mathbf{g}_{\varepsilon_k}^*)$$

710 and define the following differences:

$$\begin{aligned} p_k &:= \text{Proj}_{\mathcal{C}}(\mathbf{g}_k - \gamma_{k+1} \nabla_{\mathbf{g}} h_{\varepsilon_k}(\mathbf{g}_k, X_{k+1})) - (\mathbf{g}_k - \gamma_{k+1} \nabla_{\mathbf{g}} h_{\varepsilon_k}(\mathbf{g}_k, X_{k+1})), \\ \xi_{k+1} &:= \nabla H_{\varepsilon_k}(\mathbf{g}_k) - \nabla_{\mathbf{g}} h_{\varepsilon_k}(\mathbf{g}_k, X_{k+1}), \\ \sigma_k &:= \nabla H_0(\mathbf{g}_k) - \nabla H_{\varepsilon_k}(\mathbf{g}_k) \\ \delta_k &:= \nabla H_0(\mathbf{g}_k) - \nabla_*^2(\mathbf{g}_k - \mathbf{g}_0^*). \end{aligned}$$

711 The term p_k represents the difference between the projected and non-projected steps. Note that
 712 $p_k = 0$ if $\mathbf{g}_k - \gamma_{k+1} \nabla_{\mathbf{g}} h_{\varepsilon_k}(\mathbf{g}_k, X_{k+1}) \in \mathcal{C}$. The term ξ_k is a martingale difference ξ_k representing
 713 the difference between the regularized gradient and its non-biased estimator. σ_k represents the
 714 difference between the ε_k -regularized gradient and the non-regularized gradient. Finally, δ_k represents
 715 the difference between the gradient at \mathbf{g}_k and its linear approximation given by the Hessian at the
 716 optimum.

717 Let I_M denote identity matrix of $\mathcal{M}_M(\mathbb{R})$, observe that for any $k \in \mathbb{N}$

$$\begin{aligned} \mathbf{g}_{k+1} - \mathbf{g}_0^* &= \text{Proj}_{\mathcal{C}}(\mathbf{g}_k - \gamma_{k+1} \nabla_{\mathbf{g}} h_{\varepsilon_k}(\mathbf{g}_k, X_{k+1})) - \mathbf{g}_0^* \\ &= \mathbf{g}_k - \gamma_{k+1} \nabla_{\mathbf{g}} h_{\varepsilon_k}(\mathbf{g}_k, X_{k+1}) - \mathbf{g}_0^* - p_k && \text{incorporating } p_k \\ &= \mathbf{g}_k - \gamma_{k+1} \nabla H_{\varepsilon_k}(\mathbf{g}_k) - \mathbf{g}_0^* + \gamma_{k+1} \xi_{k+1} - p_k && \text{incorporating } \xi_{k+1} \\ &= \mathbf{g}_k - \gamma_{k+1} \nabla H_0(\mathbf{g}_k) + \gamma_{k+1} \sigma_k - \mathbf{g}_0^* + \gamma_{k+1} \xi_{k+1} - p_k && \text{incorporating } \sigma_k \\ &= (I_M - \gamma_{k+1} \nabla_k^2)(\mathbf{g}_k - \mathbf{g}_0^*) - \gamma_{k+1} \delta_k + \gamma_{k+1} \sigma_k + \gamma_{k+1} \xi_{k+1} + p_k. && \text{incorporating } \delta_k \end{aligned}$$

718 Thus, we have that

$$\nabla_*^2(\mathbf{g}_k - \mathbf{g}_0^*) = \frac{\mathbf{g}_k - \mathbf{g}_{k+1}}{\gamma_{k+1}} - \delta_k + \sigma_k + \xi_{k+1} + \frac{p_k}{\gamma_{k+1}}.$$

719 Observe that there is an orthogonal matrix U such that $\nabla_*^2 = U \text{diag}(\lambda_1, \dots, \lambda_{M-1}, 0) U^\top$. There-
 720 fore, in the following, we denote

$$(\nabla^2)^{-1} = U \text{diag}(\lambda_1^{-1}, \dots, \lambda_{M-1}^{-1}, 0) U^\top$$

721 the inverse of ∇_*^2 , restricted to the subspace $\text{Vect}(\mathbf{1}_M)^\perp$. Note that we have [18, Theorem 3.2]

$$\min_{j \in \llbracket 1, M-1 \rrbracket} \lambda_j \geq \rho_*, \quad \text{for all } k \geq 0.$$

722 Taking all the equalities in $\text{Vect}(\mathbf{1}_M)^\perp$, that is, considering all our vectors in the subspace
 723 $\text{Vect}(\mathbf{1}_M)^\perp$, we have

$$\begin{aligned} (\bar{\mathbf{g}}_t - \mathbf{g}_0^*) &= \frac{1}{t+1} \sum_{k=0}^t (\nabla_*^2)^{-1} \frac{\mathbf{g}_k - \mathbf{g}_{k+1}}{\gamma_{k+1}} - \frac{1}{t+1} \sum_{k=0}^t (\nabla_*^2)^{-1} \delta_k \\ &\quad + \frac{1}{t+1} \sum_{k=0}^t \sigma_k + \frac{1}{t+1} \sum_{k=0}^t (\nabla_*^2)^{-1} \xi_{k+1} + \frac{1}{t+1} \sum_{k=0}^t (\nabla_*^2)^{-1} \frac{p_k}{\gamma_{k+1}}. \end{aligned}$$

724 We will now give the convergence rate for each sum. Note that thanks to the introduction of σ_k , we
 725 will directly be able to use the local smoothness and strong convexity of H_0 , proved in our setting in
 726 [32].

727 • **Convergence rate for** $\frac{1}{t+1} \sum_{k=0}^t \frac{\mathbf{g}_k - \mathbf{g}_{k+1}}{\gamma_{k+1}}$.

$$\begin{aligned} \sum_{k=0}^t \frac{\mathbf{g}_k - \mathbf{g}_{k+1}}{\gamma_{k+1}} &= \sum_{k=0}^t \frac{(\mathbf{g}_k - \mathbf{g}^*) - (\mathbf{g}_{k+1} - \mathbf{g}^*)}{\gamma_{k+1}} \\ &= \sum_{k=0}^t \frac{\mathbf{g}_k - \mathbf{g}^*}{\gamma_{k+1}} - \sum_{k=0}^t \frac{\mathbf{g}_{k+1} - \mathbf{g}^*}{\gamma_{k+1}} \\ &= \sum_{k=1}^t \left(\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k} \right) (\mathbf{g}_k - \mathbf{g}^*) + \frac{\mathbf{g}_0 - \mathbf{g}^*}{\gamma_1} - \frac{\mathbf{g}_{t+1} - \mathbf{g}^*}{\gamma_{t+1}}. \end{aligned}$$

728 Remark that $\gamma_{t+1}^{-1} - \gamma_t^{-1} \leq 2\gamma_1^{-1}n^{b-1}$. By Theorem 1 (non-averaged iterates), $\mathbb{E} [\|\mathbf{g}_n - \mathbf{g}^*\|_v^2] \lesssim$
 729 $\frac{\gamma_1}{\rho_*}(t+1)^{-b}$. Therefore

$$\mathbb{E} \left[\left\| \sum_{k=0}^t \frac{\mathbf{g}_k - \mathbf{g}_{k+1}}{\gamma_{k+1}} \right\|_v^2 \right]^{\frac{1}{2}} \lesssim \frac{1}{\rho_*} \Psi_{1-b/2}(t+1) + D_C \gamma_1^{-1} + \frac{1}{\sqrt{\gamma_1 \rho_*}} (t+1)^{b/2}.$$

730 We thus have the convergence rate

$$\frac{1}{t+1} \mathbb{E} \left[\left\| \sum_{k=0}^t \frac{\mathbf{g}_k - \mathbf{g}_{k+1}}{\gamma_{k+1}} \right\|_v^2 \right]^{\frac{1}{2}} \lesssim \frac{1}{\rho_* (t+1)^{1-b/2}}.$$

731 • **Convergence rate for** $\frac{1}{t+1} \sum_{k=0}^t \delta_k$.

732 By [32, Theorem 1.3], there exists a ball $B(\mathbf{g}^*, d_1)$ with $d_1 > 0$ where H is α -Hölder. Therefore, by
 733 applying a Taylor expansion of $\nabla H_0(\mathbf{g})$ around \mathbf{g}^* , if $\mathbf{g}_k \in B(\mathbf{g}^*, d_1)$, we have

$$\|\delta_k\| \lesssim \|\mathbf{g}_k - \mathbf{g}^*\|_v^{1+\alpha}.$$

734 Otherwise, since the Hessian H_0 is uniformly bounded [32, Theorem 1.1], there exists a constant C_δ
 735 such that for any $\mathbf{g} \in \mathcal{C}$, $\|\nabla H(\mathbf{g}) - \nabla^2 H(\mathbf{g}^*)(\mathbf{g} - \mathbf{g}^*)\| \leq C_\delta$.

736 Since $\mathbb{P}(\mathbf{g}_k \notin B(\mathbf{g}^*, d_1)) = \mathbb{P}(\|\mathbf{g}_k - \mathbf{g}^*\| > d_1)$, we obtain by Markov's inequality

$$\begin{aligned} \mathbb{E}[\|\delta_k\|_v^2] &= \mathbb{E}[\|\delta_k\|_v^2 \mathbf{1}_{\mathbf{g}_k \in B(\mathbf{g}^*, d_1)}] + \mathbb{E}[\|\delta_k\|_v^2 \mathbf{1}_{\mathbf{g}_k \notin B(\mathbf{g}^*, d_1)}] \\ &\lesssim \mathbb{E}[\|\mathbf{g}_k - \mathbf{g}^*\|_v^{2+2\alpha}] + \frac{C_\delta^2}{d_1^{2+2\alpha}} \mathbb{E}[\|\mathbf{g}_n - \mathbf{g}^*\|_v^{2+2\alpha}] \\ &\lesssim \mathbb{E}[\|\mathbf{g}_k - \mathbf{g}^*\|_v^{2+2\alpha}]. \end{aligned}$$

737 Therefore, using Minkowski's inequality, we have

$$\begin{aligned} \frac{1}{t+1} \mathbb{E} \left[\left\| \sum_{k=0}^t \delta_k \right\|_v^2 \right]^{\frac{1}{2}} &\lesssim \frac{1}{t+1} \sum_{k=0}^t \frac{1}{\rho_*^{\frac{1+\alpha}{2}}} \gamma_{k+1}^{\frac{1+\alpha}{2}} \\ &\leq \frac{1}{\rho_*^{\frac{1+\alpha}{2}} (t+1)} \Psi_{\frac{b+\alpha b}{2}} \\ &\lesssim \frac{1}{\rho_*^{\frac{1+\alpha}{2}} (t+1)^{\frac{b+\alpha b}{2}}}. \end{aligned}$$

738 • **Convergence rate for** $\frac{1}{t+1} \sum_{k=0}^t \xi_{k+1}$.

739 We recall that $\xi_{k+1} = \nabla H(\mathbf{g}_k) - \nabla_{\mathbf{g}} h(\mathbf{g}_k, X_{k+1})$ and thus $\mathbb{E}[\xi_{k+1}] = 0$.

740 Observe that $\mathbb{E} \left[\left\langle \sum_{k=0}^{n-1} \xi_{k+1}, \xi_{t+1} \right\rangle_v \right] = \mathbb{E} \left[\left\langle \sum_{k=0}^{n-1} \xi_{k+1}, \mathbb{E}[\xi_{t+1} | \mathcal{F}_t] \right\rangle_v \right] = 0$.

741 Thus, since $\mathbb{E}[\|\xi_k\|^2] \leq 4$ for all k , we have the convergence rate

$$\frac{1}{t+1} \mathbb{E} \left[\left\| \sum_{k=0}^t \xi_{k+1} \right\|_v^2 \right]^{\frac{1}{2}} \leq \frac{2}{\sqrt{t+1}}.$$

742 • **Convergence rate of** $\frac{1}{t+1} \sum_{k=0}^t \sigma_k$.

743 Using Proposition 4, we have uniformly in $\mathbf{g}_k \in \mathcal{C}$ that, for all $\alpha' \in (0, \alpha)$,

$$\|\sigma_k\| = \|\nabla H_0(\mathbf{g}_k) - \nabla H_{\varepsilon_k}(\mathbf{g}_k)\| \lesssim \varepsilon^{1+\alpha'} \lesssim t^{a+\alpha'}.$$

744 Therefore

$$\begin{aligned} \frac{1}{t+1} \left\| \sum_{k=0}^t \sigma_k \right\| &\lesssim \frac{1}{t+1} \Psi_{a+\alpha'}(t) \\ &\lesssim \frac{1}{t^{a+\alpha'}}. \end{aligned}$$

745 • **Convergence rate for** $\frac{1}{t+1} \sum_{k=0}^t \frac{p_k}{\gamma_k}$.

746 Take d_0 such that $B(\mathbf{g}^*, d_0) \subset \mathcal{C}$. Defining $\nabla_k := \nabla_{\mathbf{g}} h(\mathbf{g}_k, X_{k+1})$ for conciseness, we obtain

$$\begin{aligned} \mathbb{E}[\|p_k\|_v^2] &= \mathbb{E}[\|\text{Proj}_{\mathcal{C}}(\mathbf{g}_k - \gamma_{k+1} \nabla_k) - (\mathbf{g}_k - \gamma_{k+1} \nabla_k)\|_v^2] \\ &= \mathbb{E}[\|\text{Proj}_{\mathcal{C}}(\mathbf{g}_k - \gamma_{k+1} \nabla_k) - (\mathbf{g}_k - \gamma_{k+1} \nabla_k)\|_v^2 \mathbf{1}_{\mathbf{g}_k - \gamma_{k+1} \nabla_k \notin \mathcal{C}}] \end{aligned}$$

747 Since for any $y \in \mathcal{C}$, one has $\|x - \text{Proj}_{\mathcal{C}}(x)\|_v \leq \|x - y\|_v$, taking $y = \mathbf{g}_k$, and since $\mathbf{g}_k - \gamma_{k+1} \nabla_k \notin \mathcal{C}$
748 is satisfied only if $\|\mathbf{g}_k - \gamma_{k+1} \nabla_k - \mathbf{g}^*\|_v > d_0$, we have

$$\begin{aligned} \mathbb{E}[\|p_k\|_v^2] &\leq \mathbb{E}[\|\gamma_{k+1} \nabla_k\|_v^2 \mathbf{1}_{\|\mathbf{g}_k - \gamma_{k+1} \nabla_k - \mathbf{g}^*\|_v > d_0}] \\ &\leq 4\gamma_{k+1}^2 \frac{\mathbb{E}[\|\mathbf{g}_k - \gamma_{k+1} \nabla_k - \mathbf{g}^*\|_v^4]}{d_0^4} \\ &\leq \frac{\gamma_{k+1}^2}{d_0^4} \left(2^5 \mathbb{E}[\|\mathbf{g}_k - \mathbf{g}^*\|_v^4] + 2^9 \gamma_{k+1}^4 \right) \\ &\lesssim \frac{1}{\rho_*^2} \gamma_{k+1}^4 \end{aligned}$$

749 We thus have

$$\begin{aligned} \frac{1}{t+1} \mathbb{E} \left[\left\| \sum_{k=0}^t \frac{p_k}{\gamma_k} \right\|_v^2 \right]^{\frac{1}{2}} &\lesssim \frac{1}{t+1} \sum_{k=0}^t \frac{\gamma_{k+1}}{\rho_*} \\ &\lesssim \frac{\gamma_1}{\rho_*(t+1)^b}. \end{aligned}$$

750 • **Conclusion.**

751 Finally, summing up all the convergence rates, using Cauchy-Schwarz inequality and that $(A+B)^2 \leq$
752 $2(A^2 + B^2)$ for any $A, B \in \mathbb{R}$ we obtain

$$\mathbb{E}[\|\nabla^2 H(\mathbf{g}^*)(\bar{\mathbf{g}}_n - \mathbf{g}^*)\|_v^2] \lesssim \frac{1}{\rho_*^2(t+1)^{2-b}} + \frac{1}{t^{2a+2a\alpha'}} + \frac{1}{\rho_*^{\frac{1+\alpha}{\alpha}}(t+1)^{b+\alpha b}} + \frac{1}{t+1} + \frac{\gamma_1^4}{\rho_*^2(t+1)^{2b}}.$$

Since $b > 2a$ and $b + \alpha a > 2a + 2a\alpha'$ and so noting $s = \min\{1, 2a + 2a\alpha'\}$, and since the Hessian norm is uniformly bounded, we finally obtain

$$\mathbb{E} \left[\|\bar{\mathbf{g}}_n - \mathbf{g}^*\|_v^2 \right] \lesssim \frac{1}{ts}.$$

□

B.3 Proof of Theorem 3: Convergence of the OT map estimator

Proof. We show that the convergence rate of $\bar{\mathbf{g}}_t$ to \mathbf{g}_0^* implies a convergence rate a convergence rate for the map estimation. The Brenier map is given by $T_{\mu, \nu}(x) = x - \nabla(\mathbf{g}_0^*)^c(x)$; see for instance [46], Theorem 1.17. We thus focus on the convergence of $\nabla \bar{\mathbf{g}}_t^c$ to $\nabla(\mathbf{g}_0^*)^c$.

For all $j \in \llbracket 1, M \rrbracket$, if x lies in the interior of $\mathbb{L}_j(\mathbf{g})$, we have

$$\nabla \mathbf{g}^c(x) = x - y_j. \quad (16)$$

Therefore, given $\mathbf{g}, \mathbf{g}' \in \mathbb{R}^M$, if there exists a $j \in \llbracket 1, M \rrbracket$ such that x is the interior of $\mathbb{L}_j(\mathbf{g}) \cap \mathbb{L}_j(\mathbf{g}')$ we have

$$\nabla \mathbf{g}^c(x) = \nabla(\mathbf{g}')^c(x).$$

We will now follow arguments from [46], Section 6.4.2. Fix $j, j' \in \llbracket 1, M \rrbracket$ such that $j \neq j'$ and x is in the interior of $\mathbb{L}_j(\mathbf{g}) \cap \mathbb{L}_{j'}(\mathbf{g}')$. By definition of the c -transform, we observe that $\mathbb{L}_j(\mathbf{g})$ is defined by $M - 1$ linear inequalities of the form

$$\langle x, y_{j'} - y_j \rangle \leq a_{\mathbf{g}}(j, j') := g_j - g_{j'} + \frac{1}{2}\|y_{j'}\|_2^2 - \frac{1}{2}\|y_j\|_2^2.$$

Similarly, interchanging the role of \mathbf{g}, \mathbf{g}' and j, j' we have

$$\langle x, y_j - y_{j'} \rangle \leq a_{\mathbf{g}'}(j', j) := g'_{j'} - g'_j + \frac{1}{2}\|y_j\|_2^2 - \frac{1}{2}\|y_{j'}\|_2^2.$$

We obtain that

$$\mathbb{L}_j(\mathbf{g}) \cap \mathbb{L}_{j'}(\mathbf{g}') \subset \{x \in \mathbb{R}^d : -a_{\mathbf{g}'}(j', j) \leq \langle x, y_{j'} - y_j \rangle \leq a_{\mathbf{g}}(j, j')\}.$$

Moreover, noting $h = (h_1, \dots, h_M) = \mathbf{g} - \mathbf{g}'$, we see that

$$|a_{\mathbf{g}'}(j', j) + a_{\mathbf{g}}(j, j')| \leq |h_{j'} - h_j|. \quad (17)$$

We have

$$\begin{aligned} \mu(\mathcal{A} &:= \{x \in \mathbb{R}^d, \nabla \mathbf{g}^c(x) \neq \nabla(\mathbf{g}')^c(x)\}) \\ &= \mu\left(\bigcup_{j < j'} \mathbb{L}_j(\mathbf{g}) \cap \mathbb{L}_{j'}(\mathbf{g}')\right) \\ &\leq \sum_{j < j'} \mu(\mathbb{L}_j(\mathbf{g}) \cap \mathbb{L}_{j'}(\mathbf{g}')) \\ &\leq \sum_{j < j'} \mu(\{x \in \mathbb{R}^d : -a_{\mathbf{g}'}(j', j) \leq \langle x, y_{j'} - y_j \rangle \leq a_{\mathbf{g}}(j, j')\}) . \end{aligned}$$

Under Assumption 1, μ is a measure such that $\text{Supp}(\mu) \subset B(0, R)$ and it admits a density $d\mu$ bounded by $d\mu_{\max}$. Thus,

$$\begin{aligned} \mu(\mathcal{A}) &\leq d\mu_{\max} \sum_{j < j'} \lambda_{\mathbb{R}^d}(\{x \in B(0, R) : -a_{\mathbf{g}'}(j', j) \leq \langle x, y_{j'} - y_j \rangle \leq a_{\mathbf{g}}(j, j')\}) \\ &\leq d\mu_{\max} \sum_{j < j'} \lambda_{\mathbb{R}^d}\left(\left\{x \in B(0, R) : -\frac{a_{\mathbf{g}'}(j', j)}{\|y_{j'} - y_j\|_2} \leq \left\langle x, \frac{y_{j'} - y_j}{\|y_{j'} - y_j\|_2} \right\rangle \leq \frac{a_{\mathbf{g}}(j, j')}{\|y_{j'} - y_j\|_2} \right\}\right) \\ &\leq d\mu_{\max} \sum_{j < j'} \lambda_{\mathbb{R}^d}\left(\left\{x \in B(0, R) : -\frac{a_{\mathbf{g}'}(j', j)}{\|y_{j'} - y_j\|_2} \leq x_1 \leq \frac{a_{\mathbf{g}}(j, j')}{\|y_{j'} - y_j\|_2} \right\}\right), \end{aligned}$$

by the rotational invariance of the Lebesgue measure. Combining this remark with (17) yields

$$\mu(\mathcal{A}) \leq d\mu_{\max} R^{d-1} \sum_{j < j'} \frac{|h_{j'} - h_j|}{\|y_{j'} - y_j\|_2}.$$

Similarly, for the L^p norm of the map difference, we obtain

$$\begin{aligned} \|\| (\nabla \mathbf{g}^c(\cdot) - \nabla (\mathbf{g}')^c(\cdot)) \|_q \|_{L^p(\mu)}^p &\leq \sum_{j < j'} \int_{\mathbb{L}_j(\mathbf{g}) \cap \mathbb{L}_{j'}(\mathbf{g}')} \| (\nabla \mathbf{g}^c(\cdot) - \nabla (\mathbf{g}')^c(\cdot)) \|_q d\mu(x) \\ &\leq \sum_{j < j'} \|y_{j'} - y_j\|_q \mu(\mathbb{L}_j(\mathbf{g}) \cap \mathbb{L}_{j'}(\mathbf{g}')) \\ &\leq d\mu_{\max} R^{d-1} \sum_{j < j'} \frac{\|y_{j'} - y_j\|_q |h_{j'} - h_j|}{\|y_{j'} - y_j\|_2} \\ &\leq d\mu_{\max} M^{(2-q)+/2q} R^{d-1} 2M \|h\|_1. \end{aligned}$$

So, in particular, there exists a constant $C_\Delta > 0$, independent of the location of the points y_j , which grows at least linearly in M such that

$$\|\| (\nabla \mathbf{g}^c(\cdot) - \nabla (\mathbf{g}')^c(\cdot)) \|_q \|_{L^p(\mu)}^p \leq C_\Delta \|\mathbf{g} - \mathbf{g}'\|_1 \leq C_\Delta \sqrt{M} \|\mathbf{g} - \mathbf{g}'\|.$$

Plugging in the convergence rate of $\bar{\mathbf{g}}_t$ to \mathbf{g}^* concludes the proof. \square

B.4 Proof of Corollary 1: OT cost estimation

Proof. For any vector $\mathbf{g} \in \mathbb{R}^M$, we recall the definition of $\mathbb{L}(\mathbf{g}) = \bigcup_{j=1}^M \mathbb{L}_j(\mathbf{g})$:

$$\text{for all } j \in \llbracket 1, M \rrbracket, \mathbb{L}_j(\mathbf{g}) := \left\{ x \in \mathbb{R}^d; \mathbf{g}^c(x) = \frac{1}{2} \|x - y_j\|_2^2 - g_j \right\}.$$

Note that $\mathbb{L}(\mathbf{g})$ defines a partition of \mathbb{R}^d up to μ -null sets, i.e. $\mu(\mathbb{L}_i(\mathbf{g}) \cap \mathbb{L}_j(\mathbf{g})) = 0$ when $i \neq j$, and the convex sets $\mathbb{L}_j(\mathbf{g})$ are called power or Laguerre cells. We define the set

$$\mathcal{K}^\delta := \{ \mathbf{g} : \mathbb{R}^M \rightarrow \mathbb{R} \mid \forall i \in \llbracket 1, M \rrbracket, \mu(\mathbb{L}_i(\mathbf{g})) > \delta \}.$$

Using Theorem 4.1 in [32], under Assumption 1, H_0 is uniformly $C^{2,\alpha}$ on \mathcal{K}^δ . That is, there exists a constant L such that H_0 is L -smooth on \mathcal{K}^δ . Note that the constant L depends on μ_{\min} , δ , R . We refer to [32], Remark 4.1 for more details.

By the first order condition, as soon as $\delta \leq w_{\min}$, we have $\mathbf{g}^* \in \mathcal{K}^\delta$. Indeed, at the optimum, we have for all $i \in \llbracket 1, M \rrbracket$, $\mu(\mathbb{L}_i(\mathbf{g}^*)) = w_i$. We fix here $\delta = \frac{1}{10} w_{\min}$.

Thanks to the L -smoothness, for any $\mathbf{g} \in \mathcal{K}^\delta$, we have

$$|H_0(\mathbf{g}) - H_0(\mathbf{g}^*)| \leq \frac{L}{2} \|\mathbf{g} - \mathbf{g}^*\|^2.$$

Note that, for any $\mathbf{g} \in \mathbb{R}^M$ and $i \in \llbracket 1, M \rrbracket$, the difference of measure of the Laguerre cells $\mathbb{L}_i(\mathbf{g})$ and $\mathbb{L}_i(\mathbf{g}^*)$ is at most linear with respect to $\|\mathbf{g} - \mathbf{g}^*\|_\infty$. We refer to Theorem 3 or Section 6.4.2 in [46] for more details.

Therefore, there exists a constant C_L such that, as soon as $\|\mathbf{g} - \mathbf{g}^*\|^2 \leq C_L$, we have that $\mathbf{g} \in \mathcal{K}^\delta$. This constant depends on δ, μ_{\max} , R and d as in Theorem 3. Using Theorem 2, $\mathbb{E}[\|\bar{\mathbf{g}}_t - \mathbf{g}^*\|^2] = \mathcal{O}(t^{-s})$ with $s > 0$. Then

$$\begin{aligned} \mathbb{E}[\|H_0(\bar{\mathbf{g}}_t) - H_0(\mathbf{g}^*)\|] &= \mathbb{E}[|H_0(\bar{\mathbf{g}}_t) - H_0(\mathbf{g}^*)| \mathbb{1}_{\bar{\mathbf{g}}_t \in \mathcal{K}^\delta}] + \mathbb{E}[|H_0(\bar{\mathbf{g}}_t) - H_0(\mathbf{g}^*)| \mathbb{1}_{\bar{\mathbf{g}}_t \notin \mathcal{K}^\delta}] \\ &\leq \frac{L}{2} \mathbb{E}[\|\bar{\mathbf{g}}_t - \mathbf{g}^*\|^2] + \max_{\mathbf{g} \in \mathcal{C}} |H_0(\mathbf{g}) - H_0(\mathbf{g}^*)| \mathbb{E}[\mathbb{1}_{\bar{\mathbf{g}}_t \notin \mathcal{K}^\delta}] \\ &\leq \frac{L}{2} \mathbb{E}[\|\bar{\mathbf{g}}_t - \mathbf{g}^*\|^2] + \max_{\mathbf{g} \in \mathcal{C}} |H_0(\mathbf{g}) - H_0(\mathbf{g}^*)| \mathbb{E}[\mathbb{1}_{\|\bar{\mathbf{g}}_t - \mathbf{g}^*\|^2 > C_L}] \\ &= \mathcal{O}(\mathbb{E}[\|\bar{\mathbf{g}}_t - \mathbf{g}^*\|^2]), \end{aligned}$$

where the Markov inequality of order 1 was used on $\mathbb{E}[\mathbb{1}_{\|\bar{\mathbf{g}}_t - \mathbf{g}^*\|^2 > C_L}]$. \square

794 **B.5 Proof of Proposition 2: High probability being in $B(\mathbf{g}_{\varepsilon_t}^*, \varepsilon_t)$.**

795 *Proof.* The proof of this proposition relies heavily on the technical Lemma 4, which we state and
796 prove immediately after this proof.

797 We start with a base case at $\delta = u_0 = 0$, which provides an initial convergence rate for $\mathbb{E}[\Delta_t^p]$. Then,
798 by an inductive argument, we gradually increase u_n to improve this rate till the limit when n tends to
799 infinity, namely $\min\{b - a, a\}$.

800 **Base case ($u_0 = 0$).** Using Lemma 4, with $\lambda_{c,t} = \rho_* \frac{1 - e^{-4C_\infty}}{4C_\infty} \varepsilon_t$ if $c = 0$, and $\lambda_{c,t} = \rho_*(1 -$
801 $e^{-1})\varepsilon_t t^c$ if $c \in (0, a]$, we have

$$\mathbb{E}[\Delta_{t+1}^p] \leq \mathbb{E}[\Delta_t^p] \left(1 - \gamma_{t+1} \lambda_{0,t} + C_{1,p} \gamma_{t+1}^2\right) + C_{2,p} \lambda_{c,t}^{-p+1} \gamma_{t+1}^{p+1}.$$

802 By applying Proposition 5 and Corollary 2, we obtain the following baseline convergence rate, for all
803 $p > 0$:

$$\mathbb{E}[\Delta_t^p] \lesssim \frac{\gamma_t^p}{\varepsilon_t^p}.$$

804 **Inductive step (improving the rate).** Suppose that for some $u_n \in [0, \min\{b - a, a\})$, we already
805 have

$$\mathbb{E}[\Delta_t^p] \lesssim t^{-p(b-a+u_n)}.$$

806 Choose $c < \frac{b-a+u_n}{2}$ and set $d = b - a + u_n - 2c > 0$, which is positive by construction. By
807 Markov's inequality, we then get for all $q > 0$

$$\mathbb{P}[\Delta_t \geq t^{-2c}] = \mathbb{P}[\Delta_t^q \geq t^{-2qc}] \lesssim t^{-q(b-a+u_n-2c)} = t^{-dq}.$$

808 We take q chosen large enough so that $dq > p + 1$.

809 Consequently, applying Lemma 4,

$$\mathbb{E}[\Delta_t^p] \leq \mathbb{E}[\Delta_t^p] \left(1 - \gamma_{t+1} \lambda_{c,t} + C_{1,p} \gamma_{t+1}^2\right) + C_{2,p} \lambda_{c,t}^{-p+1} \gamma_{t+1}^{p+1} + o(\gamma_{t+1}^{p+1}).$$

810 Therefore, if we pick any $u_{n+1} \in (0, \frac{b-a+u_n}{2})$, applying Proposition 5 and Corollary 2, we see that

$$\mathbb{E}[\Delta_t^p] \lesssim \frac{\gamma_t^p}{\varepsilon_t^p} t^{-cp} \lesssim t^{-pu_{n+1}}.$$

811 As soon as $b - a > u_n$, we have $(b - a + u_n)/2 > u_n$ as a valid range upper range for u_{n+1} , so we
812 can take $u_{n+1} > u_n$ and strictly improve our convergence rate.

813 **Achievability for all $\delta \in [0, \min\{b - a, a\})$.** Finally, note that the sequence defined by $u_0 = 0$
814 and $u_{n+1} = \frac{b-a+u_n}{2}$ converges to $(b - a)$, showing that every value δ up to $(b - a)$ can be reached
815 through successive improvements. Since $c = a$ is the upper bound in Lemma 4, we can continue the
816 limit $\min\{b - a, a\}$, so for all $\delta \in [0, \min\{b - a, a\})$, we have

$$\mathbb{E}[\Delta_t^p] \lesssim \frac{\gamma_t^p t^\delta}{\varepsilon_t^p}.$$

817 Using Markov's inequality concludes the proof.

818 □

819 **Lemma 4.** For any $a, b > 0$, such that $1 + a + a\alpha > 2b$, there exists constants $C_{1,p}, C_{2,p}$, only
820 depending on γ_1 and p , such that defining

$$\lambda_{c,t} = \begin{cases} \rho_* \frac{1 - e^{-4C_\infty}}{4C_\infty} \varepsilon_t, & \text{if } c = 0, \\ \rho_*(1 - e^{-1})\varepsilon_t t^c, & \text{if } c \in (0, a]. \end{cases}$$

821 we have for any $c \in [0, a]$

$$\mathbb{E}[\Delta_{t+1}^p] \leq \mathbb{E}[\Delta_t^p] \left(1 - \gamma_{t+1} \lambda_{c,t} + C_{1,p} \gamma_{t+1}^2\right) + \gamma_{t+1} D_C^{2p} \mathbb{P}(\Delta_t \geq t^{-2c}) \mathbf{1}_{c \neq 0} + C_{2,p} \lambda_{c,t}^{-p+1} \gamma_{t+1}^{p+1}. \quad (18)$$

822 *Proof.* Let us fix $c \in [0, a]$. Starting from equation (13), raising to the power p gives

$$\Delta_{t+1}^p \leq (\Delta_t - 2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle + 5\gamma_{t+1}^2)^p. \quad (19)$$

823 We note $\binom{p}{i,j,k} = \frac{p!}{i!j!k!}$ and apply the trinomial expansion to obtain

$$\begin{aligned} \Delta_{t+1}^p &\leq \sum_{\substack{i,j,k \\ i+j+k=p}} \binom{p}{i,j,k} \Delta_t^i (-2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle)^j 5^k \gamma_{t+1}^{2k} \\ &\leq \Delta_t^p - 2p\Delta_t^{p-1} \gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle \\ &\quad - 2p(p-1)\Delta_t^{p-2} \gamma_{t+1}^2 \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle^2 5\gamma_{t+1}^2 \\ &\quad + \sum_{\substack{i,j,k \\ i+j+k=p \\ (i,j,k) \notin \{(p,0,0), (p-1,1,0), (p-2,1,1)\}}} \binom{p}{i,j,k} \Delta_t^i (-2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle)^j 5^k \gamma_{t+1}^{2k}. \end{aligned}$$

824 We divide the set $\mathcal{S} := \{(i, j, k) \in \mathbb{N}^3, i + j + k = p, (i, j, k) \notin (p, 0, 0), (p-1, 1, 0), (p-2, 1, 1)\}$
825 into the following partition

$$\begin{aligned} \mathcal{P}_a &:= (i, j, k) \in \{(p-2, 2, 0), (p-3, 3, 0), (p-1, 0, 1), (0, p, 0)\}, \\ \mathcal{P}_b &:= (\mathcal{S} \setminus \mathcal{P}_a) \cap \{i = 0\}, \\ \mathcal{P}_c &:= (\mathcal{S} \setminus \mathcal{P}_a) \cap \{i \neq 0\} \cap \{j \geq 4\} \cap \{k = 0\}, \\ \mathcal{P}_d &:= (\mathcal{S} \setminus \mathcal{P}_a) \cap \{i \neq 0\} \cap \{0 < j < 4\} \cap \{k \neq 0\}, \\ \mathcal{P}_e &:= (\mathcal{S} \setminus \mathcal{P}_a) \cap \{i \neq 0\} \cap \{j = 0\} \cap \{k \neq 0\}. \end{aligned}$$

826 In what follows, the constants C_k may depend on the constant γ_1 from the learning rate $\gamma_t = \gamma_1/t^b$,
827 since we will often use the crude bound $\gamma_t^{p+1+k} \leq \gamma_1^k \gamma_t^{p+1}$ for $k \in \mathbb{N}$. Note, however, that $\lambda_{c,t} \leq 1$
828 for all t , and therefore $\lambda_{c,t}^{-q+k} \leq \lambda_{c,t}^{-q}$ for all $q, k \in \mathbb{N}$, so the constants C_k will not depend on $\lambda_{c,t}$.

829 We also introduce, for all p , the constant

$$\Gamma_p = \frac{3p+1}{2p-1}.$$

830 **Case where** $(i, j, k) \in \mathcal{P}_a$.

831 If $(i, j, k) = (p-2, 2, 0)$:

$$\begin{aligned} &\binom{p}{p-2, 2, 0} \Delta_t^{p-2} (-2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle)^2 \\ &\leq 8p(p-1) \Delta_t^{p-1} \gamma_{t+1}^2 \quad \text{by Cauchy-Schwarz} \\ &\leq 8p(p-1) \left(\Delta_t^p \frac{p-1}{p} c_1^{p/(p-1)} + \gamma_{t+1}^p \frac{1}{p c_1^p} \right) \gamma_{t+1} \quad \text{by Young: } q = \frac{p}{p-1}, q' = p, c_1 > 0 \\ &\leq p \Delta_t^p \frac{\gamma_{t+1} \lambda_{c,t}}{\Gamma_p} + \gamma_{t+1}^{p+1} \frac{1}{p} \left(\frac{8\Gamma_p(p-1)}{\lambda_{c,t}} \right)^{p-1} \quad \text{taking } c_1 = \left(\frac{\lambda_{c,t}}{8\Gamma_p(p-1)} \right)^{(p-1)/p} \\ &\leq p \Delta_t^p \frac{\gamma_{t+1} \lambda_{c,t}}{\Gamma_p} + C_1 \lambda_{c,t}^{-p+1} \gamma_{t+1}^{p+1} \quad \text{defining } C_1 \text{ readily.} \end{aligned}$$

832 If $(i, j, k) = (p-3, 3, 0)$:

$$\begin{aligned}
& \binom{p}{p-3, 3, 0} \Delta_t^{p-3} (-2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle)^3 \\
& \leq 4^3 \Delta_t^{p-\frac{3}{2}} \gamma_{t+1}^3 \quad \text{by Cauchy-Schwarz} \\
& \leq 64 \left(\Delta_t^p \frac{p - \frac{3}{2} \binom{p}{p-3}^{2p/3}}{p} c_2^{p/(p-\frac{3}{2})} + \gamma_{t+1}^{4p/3} \frac{3}{2pc_2^{2p/3}} \right) \gamma_{t+1} \\
& \quad \text{by Young: } q = \frac{p}{p-3/2}, q' = \frac{2p}{3} \\
& \leq p \Delta_t^p \frac{\gamma_{t+1} \lambda_{c,t}}{\Gamma_p} + \gamma_{t+1}^{\frac{4}{3}p+1} \frac{3}{2p} \left(\frac{32\Gamma_p(p-1)(p-2) \binom{p}{p-3}}{3\lambda_{c,t}} \right)^{\frac{2p-3}{3}} \\
& \quad \text{taking } c_2 = \left(\frac{3\lambda_{c,t}}{32\Gamma_p(p-1)(p-2)} \right)^{\frac{p-\frac{3}{2}}{p}} \\
& \leq p \Delta_t^p \frac{\gamma_{t+1} \lambda_{c,t}}{\Gamma_p} + C_2 \lambda_{c,t}^{-p+1} \gamma_{t+1}^{p+1} \quad \text{since } \frac{4}{3}p+1 \geq p+2 \text{ and } \frac{2p-3}{3} \leq p-1.
\end{aligned}$$

833 If $(i, j, k) = (p-1, 0, 1)$:

$$\begin{aligned}
& \binom{p}{p-1, 0, 1} \Delta_t^{p-1} \gamma_{t+1}^2 \leq p \left(\frac{p-1}{p} c_3^{\frac{p-1}{p}} \Delta_t^p + \frac{5^p}{pc_3^p} \gamma_{t+1}^p \right) \gamma_{t+1} \quad \text{by Young: } q = \frac{p}{p-1}, q' = p \\
& \leq p \Delta_t^p \frac{\gamma_{t+1} \lambda_{c,t}}{\Gamma_p} + \left(\frac{5^p \Gamma_p}{\lambda_{c,t}} \right)^p \gamma_{t+1}^{p+1} \quad \text{taking } c_3 = \left(\frac{\lambda_{c,t}}{\Gamma_p} \right)^{\frac{p-1}{p}} \\
& \leq p \Delta_t^p \frac{\gamma_{t+1} \lambda_{c,t}}{\Gamma_p} + C_3 \lambda_{c,t}^{-p+1} \gamma_{t+1}^{p+1} \quad \text{defining } C_3 \text{ readily.}
\end{aligned}$$

834 If $(i, j, k) = (0, p, 0)$:

$$\begin{aligned}
& \binom{p}{0, p, 0} (-2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle)^p \\
& \leq 4^p \left(\frac{c_4^2}{2} \Delta_t^p + \frac{1}{2c_4^2} \gamma_{t+1}^{2p-2} \right) \gamma_{t+1} \quad \text{Cauchy-Schwarz and Young : } q = q' = 2 \\
& \leq \Delta_t^p \frac{\gamma_{t+1} \lambda_{c,t}}{\Gamma_p} + \frac{1}{2} \left(\frac{4^{2p} \Gamma_p}{4\lambda_{c,t}} \right) \gamma_{t+1}^{2p-1} \quad \text{taking } c_4 = \left(\frac{2\lambda_{c,t}}{4^p \Gamma_p} \right)^{\frac{1}{2}} \\
& \leq \Delta_t^p \frac{\gamma_{t+1} \lambda_{c,t}}{\Gamma_p} + C_4 \lambda_{c,t}^{-1} \gamma_{t+1}^{p+1} \quad \text{since } 2p-1 \geq p+1, \text{ and defining } C_4 \text{ readily.}
\end{aligned}$$

835 **Case where** $(i, j, k) \in \mathcal{P}_b$.

836 We have $j+k=p$ such that $j+2k \geq p+1$ since $k \neq 0$. Using the bound $\|\mathbf{g}_t - \mathbf{g}_{\varepsilon_t}^*\| \leq D_C$, we
837 obtain:

$$\begin{aligned}
& \sum_{(i,j,k) \in \mathcal{P}_b} \binom{p}{i, j, k} (-2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle)^j 5^k \gamma_{t+1}^{2k} \\
& \leq \sum_{(i,j,k) \in \mathcal{P}_b} \binom{p}{i, j, k} (4D)^j 5^k \gamma_{t+1}^{j+2k} \\
& \leq C_5 \gamma_{t+1}^{p+1} \quad \text{defining } C_5 \text{ readily.}
\end{aligned}$$

838 **Case where** $(i, j, k) \in \mathcal{P}_c$.

$$\begin{aligned}
& \sum_{(i,j,k) \in \mathcal{P}_c} \binom{p}{i,j} \Delta_t^i (-2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle)^j && \text{by Cauchy-Schwarz} \\
& \leq \sum_{(i,j,k) \in \mathcal{P}_c} \binom{p}{i,j} \Delta_t^{i+j/2} 4^j \gamma_{t+1}^{j-2} \gamma_{t+1}^2 && \text{by Young: } q = \frac{p}{i+j/2}, q' = \frac{2p}{j} \\
& \leq \sum_{(i,j,k) \in \mathcal{P}_c} \binom{p}{i,j} \left(\frac{i+j/2}{p} \Delta_t^p + \frac{j}{2p} \gamma_{t+1}^{2p(j-2)/j} \right) \gamma_{t+1}^2 \\
& \leq \sum_{(i,j,k) \in \mathcal{P}_c} \binom{p}{i,j} 4^j \left(\Delta_t^p \gamma_{t+1}^2 + \frac{1}{2} \gamma_{t+1}^{p+1} \right) && j \geq 4 \text{ so: } \frac{2p(j-2)}{j} + 2 \geq p+1 \\
& \leq 8^p \Delta_t^p \gamma_{t+1}^2 + 8^p \gamma_{t+1}^{p+1}.
\end{aligned}$$

839 **Case where** $(i, j, k) \in \mathcal{P}_d$.

$$\begin{aligned}
& \sum_{(i,j,k) \in \mathcal{P}_d} \binom{p}{i,j} \Delta_t^i (-2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle)^j 5^k \gamma_{t+1}^{2k} \\
& \leq \sum_{(i,j,k) \in \mathcal{P}_d} \binom{p}{i,j,k} 5^k \Delta_t^{i+j/2} 4^j \gamma_{t+1}^{j+2k} && \text{by Cauchy-Schwarz} \\
& \leq \sum_{(i,j,k) \in \mathcal{P}_d} \binom{p}{i,j,k} 4^j \left(c_6^q \frac{i+j/2}{p} \Delta_t^p + c_6^{-q'} \frac{j C_\gamma^{2p}}{2p} \gamma_{t+1}^{2p} \right) \\
& && \text{by Young: } q = \frac{p}{i+j/2}, q' = \frac{2p}{j+2k}.
\end{aligned}$$

840 Taking $c_6 = \gamma_t^{2/q}$, it comes $c_6^{-q'} = \gamma_t^{-2q'/q} = \gamma_t^{-2/(q-1)}$. Since we are only considering cases with
841 $i, j, k \geq 1$ (which forces $p \geq 3$) and we are excluding the particular case $(i, j, k) = (p-2, 1, 1)$, one
842 can show that the parameter $q = \frac{2p}{2p-2i-j} = \frac{2p}{2k+j}$ satisfies

$$\begin{aligned}
\frac{2p}{2p-4} &\leq q \leq \frac{2p}{3} \\
\frac{4}{2p-4} &\leq q-1 \leq \frac{2p-3}{3}.
\end{aligned}$$

843 Thus, since $\frac{2}{q-1} \leq p-2$, it follows that

$$2p - \frac{2}{q-1} \geq 2p - (p-2) = p+2 \geq p+1.$$

844 Therefore, using the crude bound $\sum_{(i,j,k) \in \mathcal{P}_d} \binom{p}{i,j,k} \leq 3^p$ and defining a constant C_6 readily, we
845 obtain

$$\sum_{(i,j,k) \in \mathcal{P}_d} \binom{p}{i,j} \Delta_t^i (-2\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle)^j 5^k \gamma_{t+1}^{2k} \leq 3^p \gamma_{t+1}^2 \Delta_t^p + C_6 \gamma_{t+1}^{p+1}.$$

846 **Case where** $(i, j, k) \in \mathcal{P}_e$.

847 Since $j = 0, i+k = p$, and $(p-1, 0, 1) \in \mathcal{P}_e$, we have $k \geq 2$. We use Young's inequality with
848 $q = \frac{p}{i}, q' = \frac{p}{k}$ to obtain

$$\begin{aligned}
\sum_{(i,j,k) \in \mathcal{P}_e} \binom{p}{i,k} \Delta_t^i \gamma_{t+1}^{2k} 5^k &\leq \sum_{(i,j,k) \in \mathcal{P}_e} \binom{p}{i,k} \left(\frac{i}{p} \Delta_t^p + \frac{k}{p} (\gamma_{t+1} 5^{\frac{1}{2}})^{\frac{p(2k-2)}{k}} \right) \gamma_{t+1}^2 \\
&\leq \sum_{(i,j,k) \in \mathcal{P}_e} \binom{p}{i,k} \left(\Delta_t^p \gamma_{t+1}^2 + 5^{p-\frac{p}{k}} \gamma_{t+1}^{2p-\frac{2p}{k}+2} \right) \\
&\leq 2^p \Delta_t^p \gamma_{t+1}^2 + C_7 \gamma_{t+1}^{p+1} && \text{since } 2p - \frac{2p}{k} + 2 \geq p+2.
\end{aligned}$$

849 **Summing up the inequalities**, we obtain

$$\begin{aligned}\Delta_{t+1}^p &\leq \Delta_t^p \\ &\quad - 2p\Delta_t^{p-1}\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle \\ &\quad - 2p(p-1)\Delta_t^{p-1}\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle 5\gamma_{t+1}^2 \\ &\quad + \Delta_t^p \lambda_{c,t} \frac{3p+1}{\Gamma_p} \gamma_{t+1} \\ &\quad + \Delta_t^p \gamma_{t+1}^2 (8^p + 3^p + 2^p) \\ &\quad + \gamma_{t+1}^{p+1} \left((C_1 + C_2 + C_3) \lambda_{c,t}^{-p+1} + C_4 \lambda_{c,t}^{-1} + C_5 + 8^p + C_6 + C_7 \right).\end{aligned}$$

850 By convexity of H_{ε_t} , taking the conditional expectation gives

$$\mathbb{E} \left[-2p(p-1)\Delta_t^{p-1}\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle 5\gamma_{t+1}^2 \mid \mathcal{F}_t \right] \leq 0,$$

851 Applying Lemma 2, recalling that $\lambda_{c,t} = \rho_* \frac{1-e^{-1}}{2\sqrt{2}c_\infty} \varepsilon_t$ if $c = 0$, and $\lambda_{c,t} = \rho_* \frac{1-e^{-1}}{\sqrt{2}} \varepsilon_t t^c$ if
852 $c \in (0, a]$, we have

$$\langle \nabla H_t(\mathbf{g}), \mathbf{g} - \mathbf{g}_t^* \rangle \geq \rho_* \frac{\varepsilon_t}{\sqrt{2}\|\mathbf{g} - \mathbf{g}_t^*\|_\infty} \Delta_t \geq \begin{cases} \lambda_{c,t} \Delta_t & \text{if } c = 0 \\ \lambda_{c,t} \Delta_t \mathbf{1}_{\Delta_t \leq t^{-2c}} & \text{if } c \in (0, a] \end{cases}. \quad (20)$$

853 Therefore,

$$\begin{aligned}\mathbb{E} \left[-2p\Delta_t^{p-1}\gamma_{t+1} \langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle \mid \mathcal{F}_t \right] \\ &= -2p\Delta_t^{p-1}\gamma_{t+1} \mathbb{E} [\langle \nabla_{\mathbf{g}} h_{\varepsilon_t}(\mathbf{g}_t, X_{t+1}), \mathbf{g}_t - \mathbf{g}_t^* \rangle \mid \mathcal{F}_t] \\ &= -2p\Delta_t^{p-1}\gamma_{t+1} \langle \nabla H_{\varepsilon_t}(\mathbf{g}_t), \mathbf{g}_t - \mathbf{g}_t^* \rangle && \text{by (20)} \\ &\leq -2p\lambda_{c,t}\gamma_{t+1}\Delta_t^p + \gamma_{t+1}D_C^{2p}\mathbf{1}_{\Delta_t \geq t^{-2c}}\mathbf{1}_{c \neq 0} && \text{using that } \Delta_t^p \leq D_C^{2p}.\end{aligned}$$

854 We now just have to sum up the inequalities. Fixing $\Gamma_p = \frac{3p+1}{2p-1}$ such that $-2p + \frac{3p+1}{\Gamma_p} = -1$,

855 $C_{1,p} = 8^p + 3^p + 2^p$, $C_{2,p} = 8^p + \sum_{k=1}^7 C_k$, and taking the expectation, we have the desired form

$$\mathbb{E}[\Delta_{t+1}^p] \leq \mathbb{E}[\Delta_t^p] (1 - \gamma_{t+1}\lambda_{c,t} + C_{1,p}\gamma_{t+1}^2) + \gamma_{t+1}D_C^{2p}\mathbb{E}[\mathbf{1}_{\Delta_t \geq t^{-2c}}]\mathbf{1}_{c \neq 0} + C_{2,p}\lambda_{c,t}^{-p+1}\gamma_{t+1}^{p+1}.$$

856 □

857 **B.6 Proof of Lemma 1: Projection step**

858 *Proof.* According to [40], any optimal pair of functions $(f_\varepsilon, g_\varepsilon)$ solving the dual formulation of
859 entropic OT with regularization $\varepsilon \geq 0$ satisfies the Schrödinger equations. That is, we can take for
860 all $y \in \mathbb{R}^d$, $g_\varepsilon(y) = f_\varepsilon^{c,\varepsilon}(y)$. Moreover, $\frac{1}{2}\|x - y\|^2$ is R -Lipschitz on $B(0, R)$. Therefore, since by
861 Assumption 1, we have $\text{Supp}(\mu) \subset B(0, R)$ and $\text{Supp}(\nu) \subset B(0, R)$, we can exploit the Lipschitz
862 property of our cost function on $B(0, R)$. Using that the (c, ε) -transform has the same modulus of
863 continuity as c (see Lemma 3.1 in [40]), we get, for all $y, y' \in \mathbb{R}^d$:

$$|f_\varepsilon^{c,\varepsilon}(y) - f_\varepsilon^{c,\varepsilon}(y')| \leq R\|y - y'\|.$$

864 That is, coming back to the function g , we have for all $j, j' \in \llbracket 1, M \rrbracket$:

$$|g_\varepsilon(y_j) - g_\varepsilon(y_{j'})| \leq R\|y_j - y_{j'}\|.$$

865 By writing back our dual potential as a vector, that is $\mathbf{g}^* = (g_1^*, \dots, g_M^*)$, where for all $j \in$
866 $\llbracket 1, M \rrbracket$, $g_j^* = g_\varepsilon(y_j)$, we have

$$|g_j^* - g_{j'}^*| \leq R\|y_j - y_{j'}\|.$$

867 Moreover, if \mathbf{g}^* optimizes the semi-dual H_ε , then for any $\beta \in \mathbb{R}$, the vector $\mathbf{g}^* + \beta \mathbf{1}_M$ optimizes H_ε .
868 In particular, $\mathbf{g}^* - g_\varepsilon(y_1)\mathbf{1}_M$, which we rename \mathbf{g}^* , optimizes the semi-dual, with $g_1^* = 0$. Hence,
869 for all $j \in 1, \dots, M$

$$|g_{y_1}^* - g_{y_j}^*| = |g_{y_j}^*| \leq R\|y_1 - y_j\|.$$

870 That is, there exists an optimizer in the desired closed convex set. □

Remark: Note that for other costs such as $c(x, y) = \|x - y\|$ which defines the 1-Wasserstein distance, this projection set can be more relevant. Indeed, in this case, the cost is 1-Lipschitz and the projection set depends only on the target measure ν and no assumption of bounded cost is needed. In this case, the practitioner could choose the index k such that $g_k = 0$, minimizing for instance the Euclidean diameter of the corresponding set.

B.7 Proof of Lemma 2: Global and local RSC condition of H_ε

Proof. For any $\mathbf{g} \in \mathcal{C}$ and $s \in [0, 1]$, note $\mathbf{g}_s = \mathbf{g}_\varepsilon^* + s(\mathbf{g} - \mathbf{g}_\varepsilon^*)$, where \mathbf{g}_ε^* is the minimizer of H_ε satisfying $\sum_{i=1}^M g_i = \sum_{i=1}^M g_{\varepsilon,i}^*$ and define φ by

$$\varphi : s \in [0, 1] \mapsto H_\varepsilon(\mathbf{g}_s) .$$

Applying Lemma 5, whose proof is postponed until after this one, we have that

$$|\varphi'''(s)| \leq \frac{1}{\varepsilon} \varphi''(s) \max_{1 \leq j \leq M} |g_j - g_j^* - m(x, \mathbf{g}_s)| , \quad (21)$$

where for all $x \in \mathbb{R}^d$: $m(x, \mathbf{g}_s) := \sum_{j=1}^M \chi_j^\varepsilon(x, \mathbf{g}_s)(\mathbf{g}_s - \mathbf{g}_\varepsilon^*)$.

Using Hölder's inequality with the Hölder conjugates $p = 1, q = +\infty$ for δ_0 and Cauchy-Schwarz inequality as in [5] for δ_1 , we obtain

$$\frac{1}{\varepsilon} \max_{1 \leq j \leq M} |g_j - g_{\varepsilon,j}^* - m(x, \mathbf{g}_s)| \leq \begin{cases} \frac{2}{\varepsilon} \|\mathbf{g} - \mathbf{g}_\varepsilon^*\|_\infty =: \delta_0 , \\ \frac{\sqrt{2}}{\varepsilon} \|\mathbf{g} - \mathbf{g}_\varepsilon^*\| =: \delta_1 . \end{cases} \quad (22)$$

Use $\delta = \delta_0$ or $\delta_1 = 1$. Since $\sum_{i=1}^M g_i = \sum_{i=1}^M g_{\varepsilon,i}^*$, φ is strictly convex, and therefore, we can divide by $\varphi''(s)$ to obtain for $s \in [0, 1]$

$$\frac{\varphi'''(s)}{\varphi''(s)} \geq -\delta .$$

Integrating between 0 and S and using that $\int_0^S \frac{\varphi'''(s)}{\varphi''(s)} ds = \ln |\varphi''(S)| - \ln |\varphi''(0)|$ gives

$$\varphi''(s) \geq \exp(-\delta S) \varphi''(0) . \quad (23)$$

Since $\varphi''(s) = (\mathbf{g} - \mathbf{g}_\varepsilon^*)^\top \nabla^2 H_\varepsilon(\mathbf{g}_s) (\mathbf{g} - \mathbf{g}_\varepsilon^*)$, recalling that ρ^* is the second smallest eigenvalue of $\nabla^2 H_\varepsilon(\mathbf{g}_\varepsilon^*)$ gives the upper bound

$$\varphi''(0) \geq \rho^* \|\mathbf{g} - \mathbf{g}_\varepsilon^*\|^2 .$$

Then, since $\varphi'(s) = \langle \nabla H_\varepsilon(\mathbf{g}_s), \mathbf{g} - \mathbf{g}_\varepsilon^* \rangle$, an integration of (23) between 0 and 1 gives

$$\langle \nabla H_\varepsilon(\mathbf{g}), \mathbf{g} - \mathbf{g}_\varepsilon^* \rangle_v \geq \rho^* \frac{1}{\delta} (1 - \exp(-\delta)) \|\mathbf{g} - \mathbf{g}_\varepsilon^*\|^2 . \quad (24)$$

Note that the function $\delta \in (0, \infty) \mapsto \frac{1}{\delta} (1 - \exp(-\delta))$ is strictly decreasing and upper bounded by 1.

If $\varepsilon = 1$, take $\delta = \delta_0 = 2\|\mathbf{g} - \mathbf{g}_\varepsilon\|_\infty$ and use the fact that $\|\mathbf{g} - \mathbf{g}_\varepsilon\|_\infty \leq 2C_\infty$ to obtain

$$\langle \nabla H_1(\mathbf{g}), \mathbf{g} - \mathbf{g}_1^* \rangle \geq \rho^* \frac{1}{4C_\infty} (1 - e^{-4C_\infty}) \|\mathbf{g} - \mathbf{g}_\varepsilon^*\|^2 .$$

In the same way, for $\varepsilon < 1$ and $\|\mathbf{g} - \mathbf{g}_\varepsilon^*\| \leq \frac{\varepsilon}{\sqrt{2}}$, taking $\delta = \delta_1 = \sqrt{2}\|\mathbf{g} - \mathbf{g}_\varepsilon^*\| \leq 1$ we obtain

$$\langle \nabla H_\varepsilon(\mathbf{g}), \mathbf{g} - \mathbf{g}_\varepsilon^* \rangle_v \geq \rho^* (1 - e^{-1}) \|\mathbf{g} - \mathbf{g}_\varepsilon^*\|^2 ,$$

which concludes the proof. □

Lemma 5 (Helping Lemma for the RSC condition of H_ε). *For any $\mathbf{g} \in \mathcal{C}$ and $t \in [0, 1]$, define $\mathbf{g}_s = \mathbf{g}_\varepsilon^* + s(\mathbf{g} - \mathbf{g}_\varepsilon^*)$, where \mathbf{g}_ε^* is the minimizer of H_ε satisfying $\sum_{i=1}^M g_i = \sum_{i=1}^M g_{\varepsilon,i}^*$. The function φ , defined by*

$$\varphi : s \in [0, 1] \mapsto H_\varepsilon(\mathbf{g}_s) ,$$

897 satisfies

$$|\varphi'''(s)| \leq \frac{1}{\varepsilon} \varphi''(s) \max_{1 \leq j \leq M} |g_j - g_j^* - m(x, \mathbf{g}_s)| .$$

898 Where $m(x, \mathbf{g}_s) = \sum_{i=1}^M \chi_i^\varepsilon(x, \mathbf{g}_s)(g_i - g_{\varepsilon,i}^*)$.

899 *Proof.* The proof is an adaptation of the proof of Lemma A.2 in [5]. For completeness, we recall
900 all the steps of their proof that are needed for our results. Note that their recent erratum regarding
901 Lemma A.1 has no impact on Lemma A.2.

902 For any $\mathbf{g} \in \mathbb{R}^M$ and $s \in [0, 1]$, define $\mathbf{g}_s = \mathbf{g}_\varepsilon^* + s(\mathbf{g} - \mathbf{g}_\varepsilon^*)$, where \mathbf{g}_ε^* is the minimizer of H_ε
903 satisfying $\sum_{i=1}^M g_i = \sum_{i=1}^M g_{\varepsilon,i}^*$. We also define the function φ by

$$\varphi : s \in [0, 1] \mapsto H_\varepsilon(\mathbf{g}_s) .$$

904 Its first to third-order derivatives are given by

$$\begin{aligned} \varphi'(s) &= \langle \nabla H_\varepsilon(\mathbf{g}_s), \mathbf{g} - \mathbf{g}_\varepsilon^* \rangle , \\ \varphi''(s) &= (\mathbf{g} - \mathbf{g}_\varepsilon^*)^\top \nabla^2 H_\varepsilon(\mathbf{g}_s) (\mathbf{g} - \mathbf{g}_\varepsilon^*) , \\ \varphi'''(s) &= \sum_{i,j,k=1}^M \frac{\partial^3 H_\varepsilon(\mathbf{g}_s)}{\partial g_i \partial g_j \partial g_k} (\mathbf{g} - \mathbf{g}_\varepsilon^*)_i (\mathbf{g} - \mathbf{g}_\varepsilon^*)_j (\mathbf{g} - \mathbf{g}_\varepsilon^*)_k . \end{aligned}$$

905 Since for all $\mathbf{g} \in \mathbb{R}^M$, $\nabla H_\varepsilon(\mathbf{g}) = -\mathbb{E}_{X \sim \mu} [\chi^\varepsilon(X, \mathbf{g})] + \mathbf{w}$,

$$\begin{aligned} \varphi'(s) &= \langle -\mathbb{E}_{X \sim \mu} [\chi^\varepsilon(X, \mathbf{g}_s)] + \mathbf{w}, \mathbf{g} - \mathbf{g}_\varepsilon^* \rangle \\ &= -\mathbb{E}_{X \sim \mu} [m(X, \mathbf{g}_s)] + \langle \mathbf{w}, \mathbf{g} - \mathbf{g}_\varepsilon^* \rangle , \end{aligned}$$

906 defining for all $x \in \mathbb{R}^d$, $m(x, \mathbf{g}_s) = \sum_{i=1}^M \chi_i^\varepsilon(x, \mathbf{g}_s)(g_i - g_{\varepsilon,i}^*)$.

907 Using that $\nabla_{\mathbf{g}} \chi^\varepsilon(x, \mathbf{g}) = \frac{1}{\varepsilon} (\text{diag}(\chi^\varepsilon(x, \mathbf{g})) - \chi^\varepsilon(x, \mathbf{g}) \chi^\varepsilon(x, \mathbf{g})^\top)$, we have

$$\frac{d}{ds} \chi^\varepsilon(X, \mathbf{g}_s) = \frac{1}{\varepsilon} (\text{diag}(\chi^\varepsilon(X, \mathbf{g}_s)) - \chi^\varepsilon(X, \mathbf{g}_s) \chi^\varepsilon(X, \mathbf{g}_s)^\top) (\mathbf{g} - \mathbf{g}_\varepsilon^*) .$$

908 Therefore, using the expression of m yields to

$$\begin{aligned} \varphi''(s) &= -\frac{1}{\varepsilon} \mathbb{E}_{X \sim \mu} [(\mathbf{g} - \mathbf{g}_\varepsilon^*)^\top \text{diag}(\chi_i^\varepsilon(X, \mathbf{g}_s)) (\mathbf{g} - \mathbf{g}_\varepsilon^*) - m(X, \mathbf{g}_s)^2] \\ &= -\frac{1}{\varepsilon} \mathbb{E}_{X \sim \mu} [\sigma^2(X, \mathbf{g}_s)] \end{aligned}$$

909 defining for all $x \in \mathbb{R}^d$

$$\begin{aligned} \sigma^2(x, \mathbf{g}_s) &= \sum_{i=1}^M \chi_i^\varepsilon(x, \mathbf{g}_s)(g_i - g_{\varepsilon,i}^*)^2 - (m(x, \mathbf{g}_s))^2 \\ &= \sum_{i=1}^M \chi_i^\varepsilon(x, \mathbf{g}_s) (g_i - g_{\varepsilon,i}^* - m(x, \mathbf{g}_s))^2 . \end{aligned}$$

910 A derivation of σ^2 leads to (see [5] eq. (A.19),) for more details)

$$\begin{aligned} -\varepsilon \frac{d}{ds} \sigma^2(x, \mathbf{g}_s) &= \sum_{i=1}^M \chi_i(x, \mathbf{g}_s)(g_i - g_{\varepsilon,i}^*)^3 - 3m(x, \mathbf{g}_s) \sigma^2(x, \mathbf{g}_s) - (m(x, \mathbf{g}_s))^3 \\ &= \sum_{i=1}^M \chi_i^\varepsilon(x, \mathbf{g}_s) (g_i - g_{\varepsilon,i}^* - m(x, \mathbf{g}_s))^3 . \end{aligned}$$

911 Since $\varepsilon^2 \varphi'''(s) = \mathbb{E}_{X \sim \mu} [\frac{d}{ds} \sigma^2(X, \mathbf{g}_s)]$, we conclude

$$|\varphi'''(s)| \leq \frac{1}{\varepsilon} \varphi''(s) \max_{1 \leq j \leq M} |g_j - g_j^* - m(x, \mathbf{g}_s)| .$$

912

□

913 B.8 Other Technical results

914 **Proposition 4.** *For all $\mathbf{g} \in \mathcal{C}$ and all $\alpha' \in (0, \alpha)$, we have*

$$\|\nabla H_\varepsilon(\mathbf{g}) - \nabla H_0(\mathbf{g})\|_\infty \lesssim \varepsilon^{1+\alpha'}.$$

915 *Proof.* We adopt the decomposition of \mathcal{X} from Appendix A.3 of [18]. See Figure 1 in [18] for an
 916 illustration of this decomposition. Fix $i \in \{1, \dots, M\}$, let $\mathcal{X} \subset B(0, R)$ be the support of μ , and
 917 choose parameters

$$\eta = \varepsilon^\beta, \quad \gamma = \frac{1}{2}\eta, \quad \beta \in (0, 1).$$

918 Define, for all $i \in \llbracket 1, M \rrbracket$, the function

$$f_i(x) := -c(x, y_i) + g_i = -\frac{1}{2}\|x - y_i\|^2 + g_i,$$

919 and use these to define the following sets:

$$H_{ij} = \mathbb{L}_i(\mathbf{g}) \cap \mathbb{L}_j(\mathbf{g})$$

920

$$\mathcal{X}_{i,\eta,+} = \left\{ x \in \mathbb{L}_i(\mathbf{g}); \forall j \neq i, \frac{f_i(x) - f_j(x)}{\|y_i - y_j\|} \geq \eta \right\},$$

921

$$\mathcal{X}_{i,\eta,-} = \left\{ x \in \mathbb{R}^d; \arg \max_j f_j(x) = k, \frac{f_k(x) - f_i(x)}{\|y_k - y_i\|} \geq \eta \right\},$$

922

$$H_{ij}^\gamma = \{x \in H_{ij} \mid \forall k \notin \{i, j\}, f_i(x) = f_j(x) \geq f_k(x) + \gamma \max(\|y_i - y_k\|, \|y_j - y_k\|)\},$$

923

$$A_{i,\eta,\gamma} = \bigcup_{j \neq i} \{x + t d_{ij}; x \in H_{ij}^\gamma, t \in [-\eta\|y_i - y_j\|, \eta\|y_i - y_j\|]\}, \quad d_{ij} = \frac{y_i - y_j}{\|y_i - y_j\|},$$

924

$$B_{i,\eta,\gamma} = \mathbb{R}^d \setminus (\mathcal{X}_{i,\eta,+} \cup \mathcal{X}_{i,\eta,-} \cup A_{i,\eta,\gamma}).$$

925 We also recall the point-wise definitions of the regularized and non-regularized functions constituting
 926 the gradients

$$\chi_i^\varepsilon(x) = \frac{\exp(f_i(x)/\varepsilon)}{\sum_{k=1}^M \exp(f_k(x)/\varepsilon)}, \quad \chi_i(x) = \mathbf{1}\{i = \arg \max_k f_k(x)\}.$$

927 and define the constant

$$c_y = \min_{i \neq j} \|y_i - y_j\| > 0.$$

928 **Error decomposition.**

$$\|\nabla H_\varepsilon(\mathbf{g}) - \nabla H_0(\mathbf{g})\|_\infty = \max_i \left| \int_{\mathcal{X}} (\chi_i^\varepsilon - \chi_i) d\mu \right| \leq I_1 + I_2 + I_3 + I_4,$$

929 with

$$I_1 = \int_{\mathcal{X}_{i,\eta,+}} |\chi_i^\varepsilon - \chi_i| d\mu, \quad I_2 = \int_{\mathcal{X}_{i,\eta,-}} |\chi_i^\varepsilon - \chi_i| d\mu, \quad I_3 = \int_{A_{i,\eta,\gamma}} |\chi_i^\varepsilon - \chi_i| d\mu, \quad I_4 = \int_{B_{i,\eta,\gamma}} |\chi_i^\varepsilon - \chi_i| d\mu.$$

930 **1. Interior regions $\mathcal{X}_{i,\eta,+}$ and $\mathcal{X}_{i,\eta,-}$.**

931 For $x \in \mathcal{X}_{i,\eta,+}$ one has $f_i - f_j \geq \eta\|y_i - y_j\| \geq \eta c_y$ for every $j \neq i$, hence

$$\begin{aligned} |\chi_i(x) - \chi_i^\varepsilon(x)| &= 1 - \chi_i^\varepsilon(x) \\ &= 1 - \frac{e^{f_i/\varepsilon}}{\sum_{k=1}^M e^{f_k/\varepsilon}} \\ &= \frac{\sum_{k \neq i} e^{f_k/\varepsilon}}{\sum_{k=1}^M e^{f_k/\varepsilon}} \\ &\leq \mathcal{O}\left(e^{-\eta c_y/\varepsilon}\right) \end{aligned}$$

932 In a same way, we obtain the same bound if $x \in \mathcal{X}_{i,\eta,-}$, therefore

$$I_1 + I_2 = \mathcal{O}\left(e^{-\eta c_y/\varepsilon}\right).$$

933 **2. Simple slabs $A_{i,\eta,\gamma}$.**

934 Inside one slab $T_{ij} = \{x + t d_{ij} \mid x \in H_{ij}^\gamma, t \in [-\eta\|y_i - y_j\|, \eta\|y_i - y_j\|]\}$, and we have $f_i(x) -$
 935 $f_j(x) = t\|y_i - y_j\|$. for a certain $t \in [-\eta\|y_i - y_j\|, \eta\|y_i - y_j\|]$. All other indices satisfy
 936 $f_k(x) - f_i(x) \leq -c_y\gamma$, so

$$\sum_{k \notin \{i,j\}} e^{(f_k - f_i)/\varepsilon} \leq (M-2)e^{-c_y\gamma/\varepsilon}.$$

937 Hence

$$\chi_i^\varepsilon(x) = \frac{1}{1 + e^{-t\|y_i - y_j\|/\varepsilon} + \sum_{k \notin \{i,j\}} e^{(f_k(x) - f_i(x))/\varepsilon}} = p_\varepsilon(t) + \mathcal{O}\left(e^{-\gamma c_y/\varepsilon}\right),$$

938 with $p_\varepsilon(t) = (1 + e^{-t/\tilde{\varepsilon}})^{-1}$, $\tilde{\varepsilon} = \varepsilon/\|y_i - y_j\|$.

939 Introduce coordinates $x = z + t n_{ij}$ with $n_{ij} = \frac{y_i - y_j}{\|y_i - y_j\|}$ and $z \in H_{ij}$; the Jacobian of this change
 940 of coordinate is 1. Since f_μ is α -Hölder, there exists $L > 0$ such that we can write $f_\mu(z + t n_{ij}) =$
 941 $f_\mu(z) + r_\alpha(z, t)$ with $|r_\alpha(z, t)| \leq L|t|^\alpha$. Note that the function $p_\varepsilon(t) - \mathbf{1}_{t>0}$ is odd. Writing σ the
 942 Hausdorff measure of dimension $d-1$, we obtain

$$\begin{aligned} \left| \int_{T_{ij}^{\eta,\gamma}} (\chi_i^\varepsilon - \chi_i) d\mu \right| &= \left| \int_{H_{ij}^\gamma \cap B(0,R)} \int_{-\eta\|y_i - y_j\|}^{\eta\|y_i - y_j\|} [p_\varepsilon(t) - \mathbf{1}_{t>0} + \mathcal{O}(e^{-\gamma/\varepsilon})] (f_\mu(z) + r_\alpha(z, t)) dt d\sigma(z) \right| \\ &= \left| \int_{H_{ij}^\gamma \cap B(0,R)} \int_{-\eta\|y_i - y_j\|}^{\eta\|y_i - y_j\|} [p_\varepsilon(t) - \mathbf{1}_{t>0}] r_\alpha(z, t) dt d\sigma(z) \right. \\ &\quad + \int_{H_{ij}^\gamma \cap B(0,R)} \int_{-\eta\|y_i - y_j\|}^{\eta\|y_i - y_j\|} [p_\varepsilon(t) - \mathbf{1}_{t>0}] f_\mu(z) dt d\sigma(z) \\ &\quad \left. + \int_{H_{ij}^\gamma \cap B(0,R)} \int_{-\eta\|y_i - y_j\|}^{\eta\|y_i - y_j\|} \mathcal{O}(e^{-\gamma/\varepsilon}) (f_\mu(z) + r_\alpha(z, t)) dt d\sigma(z) \right| \\ &\lesssim \text{vol}_{d-1}(H_{ij}^\gamma \cap B(0,R)) \int_{-\eta\|y_i - y_j\|}^{\eta\|y_i - y_j\|} |t|^\alpha dt + \mathcal{O}\left(\eta e^{-\gamma c_y/\varepsilon}\right) \\ &= \mathcal{O}(\eta^{1+\alpha}) + \mathcal{O}\left(\eta e^{-\gamma c_y/\varepsilon}\right). \end{aligned}$$

943 Summing over $j \neq i$ yields

$$I_3 = \mathcal{O}(\eta^{1+\alpha}) + \mathcal{O}(\eta e^{-\gamma c_y/\varepsilon}).$$

944 **3. Corner set $B_{i,\eta,\gamma}$.**

945 As shown in [18], denoting by θ the maximum angle that can be formed by three non-aligned points
 946 of the target measure, each corner that constitutes $B_{i,\eta,\gamma}$ is included in a cylinder of volume at
 947 most $\frac{4\pi \text{diam}(B(0,R))^{d-2}}{\cos(\theta/2)^2} \gamma^2$. Moreover, there are at most M^2 such corners. Therefore, $\mu(B_{i,\eta,\gamma}) =$
 948 $\mathcal{O}(\gamma^2) = \mathcal{O}(\eta^2)$, and so

$$I_4 = \mathcal{O}(\eta^2).$$

949 **4. Choice of the exponent β .**

950 Let $\alpha' \in (0, \alpha)$ and pick

$$\beta \in \left(\frac{1+\alpha'}{1+\alpha}, 1\right).$$

951 Then $\eta^{1+\alpha} = \varepsilon^{\beta(1+\alpha)} \leq \varepsilon^{1+\alpha'}$ and $\eta^2 = \varepsilon^{2\beta} \leq \varepsilon^{1+\alpha'}$. Exponential terms are even smaller, hence

$$\|\nabla H_\varepsilon(\mathbf{g}) - \nabla H_0(\mathbf{g})\|_\infty = \mathcal{O}\left(\varepsilon^{1+\alpha'}\right).$$

952

□

953 **Proposition 5.** Let $(\gamma_t)_{t \geq 0}$ and $(\nu_t)_{t \geq 0}$ be some positive and decreasing sequences and let $(\delta_t)_{t \geq 0}$,
954 satisfying the following:

955 • The sequence δ_t follows the recursive relation:

$$\delta_{t+1} \leq (1 - 2\omega\gamma_{t+1} + \eta\gamma_{t+1}^2) \delta_t + \nu_{t+1}\gamma_{t+1}, \quad (25)$$

956

with $\delta_0 \geq 0$ and $\omega, \eta > 0$.

957

• Let γ_t converge to 0.

958

• Let $t_0 = \inf \{t \geq 1 : \omega\gamma_{t+1} \leq 1; \eta\gamma_t \leq \omega\}$.

959 Then, for all $t \geq t_0$, we have the upper bound:

$$\delta_t \leq \exp\left(-\omega \sum_{i=t_0+1}^t \gamma_i\right) \left(\sum_{k=t_0}^t \gamma_k \nu_k + \delta_{t_0}\right) + \frac{1}{\omega} \nu_{\lceil \frac{t}{2} \rceil - 1} \leq \frac{1}{\omega} \nu_{\lceil \frac{t}{2} \rceil - 1} + o(\nu_t).$$

960 *Proof.* For all $t \geq t_0$, since $1 - 2\omega\gamma_{t+1} + \eta\gamma_{t+1}^2 \leq 1 - \omega\gamma_{t+1}$, one has

$$\begin{aligned} \delta_t &\leq (1 - \omega\gamma_{t+1} + \eta\gamma_{t+1}^2) \delta_t + \nu_{t+1}\gamma_{t+1} \\ &\leq \underbrace{\prod_{i=t_0+1}^t (1 - \omega\gamma_i) \delta_{t_0}}_{=: U_{1,t}} + \underbrace{\sum_{k=t_0+1}^t \prod_{i=k+1}^t (1 - \omega\gamma_i) \gamma_k \nu_k}_{=: U_{2,t}} \end{aligned}$$

961 One can consider two cases: $\lceil t/2 \rceil - 1 \leq t_0$ and $\lceil t/2 \rceil - 1 > t_0$.

962 **Case where $\lceil t/2 \rceil - 1 \leq t_0 < t$:** Since ν_k is decreasing,

$$\begin{aligned} U_{2,t} &\leq \nu_{t_0+1} \sum_{k=t_0+1}^t \prod_{i=k+1}^t (1 - \omega\gamma_i) \gamma_k \\ &= \frac{1}{\omega} \nu_{t_0+1} \sum_{k=t_0+1}^t \prod_{i=k+1}^t (1 - \omega\gamma_i) - \prod_{i=k}^t (1 - \omega\gamma_i) \\ &= \frac{1}{\omega} \nu_{t_0+1} \left(1 - \prod_{i=t_0+1}^t (1 - \omega\gamma_i)\right) \\ &\leq \frac{1}{\omega} \nu_{t_0+1} \end{aligned}$$

963 Since ν_k is decreasing, it comes $U_{2,t} \leq \frac{1}{\omega} \nu_{\lceil t/2 \rceil}$.

964 **Case where $\lceil t/2 \rceil - 1 > t_0$:** As in [3], for all $m = t_0 + 1, \dots, t$, one has

$$U_{2,t} \leq \exp\left(-\omega \sum_{k=m+1}^t \gamma_k\right) \sum_{k=t_0+1}^m \gamma_k \nu_k + \frac{1}{\omega} \nu_m$$

965 Then, taking $m = \lceil t/2 \rceil - 1$, it comes

$$U_{2,t} \leq \exp\left(-\omega \sum_{k=\lceil t/2 \rceil}^t \gamma_k\right) \sum_{k=t_0+1}^{\lceil t/2 \rceil - 1} \gamma_k \nu_k + \frac{1}{\omega} \nu_{\lceil t/2 \rceil - 1}$$

966

□

967 **Corollary 2.** Let $(\gamma_t)_{t \geq 0}$ and $(\nu_t)_{t \geq 0}$ be some positive and decreasing sequences and let $(\delta_t)_{t \geq 0}$ be
 968 a sequence satisfying the following:

969 • The sequence δ_t follows the recursive relation:

$$\delta_{t+1} \leq (1 - 2\omega\gamma_{t+1} + \eta\gamma_{t+1}^2) \delta_t + \nu_{t+1}\gamma_{t+1}, \quad (26)$$

970 with $\delta_0 \geq 0$ and $\omega, \eta > 0$.

971 • Let $\gamma_t = c_\gamma t^{-\alpha}$ with $\alpha \in (0, 1)$.

972 • Let $t_0 = \inf \{t \geq 1 : \omega\gamma_{t+1} \leq 1; \eta\gamma_t \leq \omega\}$.

973 Then, for all $t \in \mathbb{N}$, we have the upper bound:

$$\delta_t \leq \frac{1}{\omega} \nu_{\frac{t}{2}-1} + o(\nu_t).$$

974 *Proof.* Applying Proposition 5, for all $t \geq t_0$, we have the upper bound:

$$\delta_t \leq \exp \left(-\omega \sum_{i=t_0+1}^t \gamma_i \right) \left(\sum_{k=t_0}^t \gamma_k \nu_k + \delta_{t_0} \right) + \frac{1}{\omega} \nu_{\lceil \frac{t}{2} \rceil - 1}.$$

975 Approximating the sum $\sum_{s=t_0}^t \gamma_s$ via a Riemann sum lower bound for the function $x \mapsto \frac{1}{x^\alpha}$, and
 976 applying the logarithmic inequality $\log(1-x) \leq -x$, one can now bound $\prod_{i=t_0+1}^t (1 - \omega\gamma_i) \delta_{t_0}$ as

$$\begin{aligned} \prod_{i=t_0+1}^t (1 - \omega\gamma_i) \delta_{t_0} &\leq \exp \left(-\omega \frac{c_\gamma}{1-\alpha} \left((t+1)^{1-\alpha} - (t_0+1)^{1-\alpha} \right) \right) \gamma_{t_0} \nu_{t_0} \\ &\leq \exp \left(-\frac{\omega c_\gamma}{2} \left((t+1)^{1-\alpha} - (t_0+1)^{1-\alpha} \right) \right) \gamma_{t_0} \nu_{t_0}. \end{aligned}$$

977 In a same way, since

$$\exp \left(-\omega \sum_{k=\lceil t/2 \rceil}^t \gamma_k \right) \leq \exp \left(-\frac{\omega c_\gamma}{2} (t+1)^{1-\alpha} \right),$$

978 we obtain

$$\delta_t \leq \exp \left(-\frac{1}{2} \omega c_\gamma t^{1-\alpha} \right) \exp \left(\frac{1}{2} \omega c_\gamma (t_0+1)^{1-\alpha} \right) \left(\sum_{k=t_0}^t \gamma_k \nu_k + \delta_{t_0} \right) + \frac{1}{\omega} \nu_{\frac{t}{2}-1}.$$

979 Since the product involving exponential terms converges exponentially fast, we finally obtain the
 980 desired convergence rate

$$\delta_t \leq \frac{1}{\omega} \nu_{\frac{t}{2}-1} + o(\nu_t).$$

981

□

982