Stochastic Optimization for Regularized Wasserstein Estimators

Marin Ballu Quentin Berthet Francis Bach

This supplementary material contains the proofs of the technical results found in the main text.

Proof of proposition 2.1. This proof follows the reasoning in (Rigollet & Weed, 2018). Let $\mu = \frac{1}{I}\sum_i \delta_{X_i}$ be the empirical measure of the sample (X_i). We first remark that the log-likelihood of X_i defined by

$$\ell_\nu(X_i) := \log \int \kappa(X_i, y) d\nu(y)$$

verifies

$$\ell_\nu(X_i) = \log \mathbb{E}_{Y\sim\nu}\left[\kappa(X_i, Y)\right].$$

With the Legendre transform of the relative entropy, we obtain

$$\ell_\nu(X_i) = \sup_{\gamma_i} \mathbb{E}_{Y\sim\gamma_i}\left[\log \kappa(X_i, Y)\right] - \mathrm{KL}(\gamma_i, \nu)$$

with the minimum being over every probability measures γ_i on $\mathcal{Y}$. The MLE maximizes

$$\frac{1}{I}\sum_i \ell_\nu(X_i) = \mathbb{E}_{X\sim\mu}\left[\ell_\nu(X)\right]$$

over $\nu \in \mathcal{M}$, it can be written as

$$\max_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\left[\log \kappa(X, Y)\right] - \mathbb{E}_{X\sim\mu}\left[KL(\pi(X,\cdot),\nu)\right],$$

with $\pi(X, \cdot)$ being the conditional probability of π, defined by $\pi(X_i, \cdot) := \gamma_i$. We have

$$\begin{aligned}
&\mathbb{E}_{X\sim\mu}\left[KL(\pi(X,\cdot),\nu)\right] \\
&= \frac{1}{I}\sum_i \mathbb{E}_{Y\sim\nu}\left[\log \frac{d\pi(X_i,\cdot)}{d\nu}(Y)\right], \\
&= \frac{1}{I}\sum_i \mathbb{E}_{Y\sim\nu}\left[\log \frac{d\pi}{d\mu\otimes\nu}(X_i, Y)\right] - \log I, \\
&= KL(\pi, \mu\otimes\nu) - \log I.
\end{aligned}$$

Thus the MLE minimizes

$$\min_{\pi\in\Pi(\mu,\nu)} \mathbb{E}\left[c(X, Y)\right] + \varepsilon \mathrm{KL}(\pi, \mu\otimes\nu),$$

which is the regularized optimal transport cost between μ and ν. □

Proof of the formulas for the gaussian case. The estimator $\hat\nu$ is a gaussian variable that minimizes

$$\begin{aligned}
&\mathrm{OT}_\varepsilon(\mu,\nu) + \eta\,\mathrm{KL}(\nu,\beta) = \\
&|m_\nu - m_\mu|^2 + \frac{\eta}{2}|m_\nu|^2 \\
&+ \sigma_\mu^2 + \left(1 + \frac{\eta}{2}\right)\sigma_\nu^2 \\
&- \sqrt{4\sigma_\mu^2\sigma_\nu^2 + \frac{\varepsilon^2}{4}} - \frac{\eta}{2}\log\sigma_\nu^2 \\
&+ \frac{\varepsilon}{2}\log\left(\varepsilon + \sqrt{4\sigma_\mu^2\sigma_\nu^2 + \frac{\varepsilon^2}{4}}\right). \qquad (1)
\end{aligned}$$

We write $\hat\nu \sim \mathcal{N}(m_{\hat\nu}, \sigma_{\hat\nu})$ The expression of $m_{\hat\nu}$ comes from the minimization of the first two terms of (1), where we take the derivative:

$$2(m_{\hat\nu} - m_\mu) - \eta m_{\hat\nu} = 0,$$

so

$$m_{\hat\nu} = \frac{m_\mu}{1 + \frac{\eta}{2}}.$$

For $\sigma_{\hat\nu}$, we note

$$\phi(x) = \sqrt{4\sigma_\mu^2 x^2 + \frac{\varepsilon^2}{4}},$$

Noting that we consider the values of $x = \sigma_\nu$ to be between 1 and σ_μ, then for $\varepsilon \to 0$ we have the Taylor expansions

$$\begin{aligned}
\phi(x) &= 2\sigma_\mu x + O(\varepsilon^2), \\
\phi'(x) &= 2\sigma_\mu + O(\varepsilon^2).
\end{aligned}$$

The derivative of (1) over σ_ν gives

$$2\left(1 + \frac{\eta}{2}\right)\sigma_{\hat\nu}^2 - \phi(\sigma_{\hat\nu}) - \frac{\eta}{\sigma_{\hat\nu}} + \frac{\varepsilon\phi'(\sigma_{\hat\nu})}{2(\varepsilon + \phi(\sigma_{\hat\nu}))} = 0.$$

We multiply by $\sigma_{\hat\nu}$ and use the Taylor expansions for $\varepsilon \to 0$,

$$2\left(1 + \frac{\eta}{2}\right)\sigma_{\hat\nu}^2 - 2\sigma_\mu\sigma_{\hat\nu} - \eta + \frac{\varepsilon}{2} = O(\varepsilon^2).$$

This second order polynomial in $\sigma_{\hat\nu}$ has two real roots, of the form

$$x = \frac{\sigma_\mu \pm \sqrt{\sigma_\mu^2 + \left(1 + \frac{\eta}{2}\right)(2\eta - \varepsilon)}}{2 + \eta},$$

one of which is negative, so $\sigma_{\hat{\nu}}$ is converging to the positive one when $\varepsilon \to 0$. Thus we have

$$\sigma_{\hat{\nu}} = \frac{\sigma_\mu + \sqrt{\sigma_\mu^2 + \left(1+\frac{\eta}{2}\right)(2\eta - \varepsilon)}}{2+\eta} + O(\varepsilon^2).$$

The second expression comes from a Taylor expansion of the square root for $\eta \to 0$. □

Proof of Proposition 3.3. 1. The function $H^*_{\beta,\mathcal{M}}$ is a Legendre transform, so it is convex, and thus $-F$ is convex as a sum of convex functions. Moreover F is bounded from above:

$$\begin{aligned} F(a,b) \leq& C_1\mathbb{E}[a_i+b_j] - C_2\mathbb{E}\left[e^{\frac{a_i+b_j}{\varepsilon}}\right], \\ \leq& C_3, \end{aligned}$$

where C_3 does not depend on a or b. Thus the set of solutions is nonempty. F is invariant by the translation $(a,b) \mapsto (a_1+c,\dots,a_I+c,b_1-c,\dots,b_J-c)$, so each solution generates an affine set of solutions spanned by the vector $((1,\dots,1),(-1,\dots,-1))$. We can conclude using the strong convexity on the slice $\{\sum_i \mu_i a_i = \sum_j \beta_j b_j\}$, which implies that there exists only one solution on this slice.

2. The solution (a^*,b^*) solves the following system

$$\begin{cases} \nabla_a F(a^*,b^*) = 0, \\ \nabla_b F(a^*,b^*) = 0. \end{cases}$$

With notations $A_i = e^{a_i^*/\varepsilon}$, $B_j = e^{b_j^*/\varepsilon}$, $\Gamma_{i,j} = e^{-C_{i,j}/\varepsilon}$, the two equations can be written as

$$\begin{cases} \forall\, 1\leq i\leq I, \quad 1 - A_i\sum_j \beta_j B_j \Gamma_{i,j} = 0, \\ \forall\, 1\leq j\leq J, \quad f_j - B_j\sum_i \mu_i A_i \Gamma_{i,j} = 0. \end{cases} \tag{2}$$

Thus

$$\begin{cases} \forall\, 1\leq i\leq I, \quad A_i = \frac{1}{\sum_j \beta_j B_j \Gamma_{i,j}}, \\ \forall\, 1\leq j\leq J, \quad B_j = \frac{f_j}{\sum_i \mu_i A_i \Gamma_{i,j}}. \end{cases} \tag{3}$$

We also remark that by multiplying the second term of (2) by β_j and summing over j we get

$$\sum_{i,j} \mu_i A_i \beta_j B_j \Gamma_{i,j} = 1. \tag{4}$$

By multiplying the equations in (3) we have for all i,j:

$$A_i B_j \Gamma_{i,j} = \frac{f_j \Gamma_{i,j}}{\sum_{k,l} \mu_k A_k \Gamma_{k,j} \beta_l B_l \Gamma_{i,l}}$$

thus using (4):

$$f_j \min_{k,l} \frac{\Gamma_{i,j}\Gamma_{k,l}}{\Gamma_{k,j}\Gamma_{i,l}} \leq A_i B_j \Gamma_{i,j} \leq f_j \max_{k,l} \frac{\Gamma_{i,j}\Gamma_{k,l}}{\Gamma_{k,j}\Gamma_{i,l}},$$

finally

$$e^{-m-2R_C/\varepsilon} \leq A_i B_j \Gamma_{i,j} \leq e^{m+2R_C/\varepsilon}.$$

3. We now prove that $-F$ is strongly convex. We compute

$$\begin{aligned} -\nabla_a^2 F &= \mathbb{E}\left[\frac{1}{\varepsilon} D_{i,j} E_{i,i}\right], \\ -\nabla_b^2 F &= -\nabla_b \nu^* + \mathbb{E}\left[\frac{1}{\varepsilon} D_{i,j} E_{j,j}\right], \\ -\nabla_a \nabla_b F &= \mathbb{E}\left[\frac{1}{\varepsilon} D_{i,j} E_{i,j}\right]. \end{aligned}$$

We remark that

$$\nu^* = \mathrm{softmax}(-b_j/\eta + \log\beta_j),$$

so

$$-\nabla_b \nu^* = \frac{1}{\eta} S$$

with

$$\begin{aligned} S &:= (\nabla \mathrm{softmax})(-b_j/\eta + \log\beta_j), \\ S &= (\nu_i(\delta_{i,j} - \nu_j))_{i,j}. \end{aligned}$$

We remark that $S \succcurlyeq 0$ since

$$\begin{aligned} u^T S u &= \sum_i \nu_i u_i^2 - \left(\sum_i \nu_i u_i\right)^2 \\ &= \mathbb{E}_\nu[U^2] - (\mathbb{E}_\nu[U])^2 \geq 0 \end{aligned}$$

by Jensen, with $U = u_j$ with probability ν_j. It implies $-\nabla_b \nu_j^* \succcurlyeq 0$. So

$$-\nabla_{a,b}^2 F \succcurlyeq \frac{1}{\varepsilon} M,$$

with

$$M := \mathbb{E}\left[D_{i,j}\begin{pmatrix} E_{i,i} & E_{i,j} \\ E_{j,i} & E_{j,j}\end{pmatrix}\right].$$

As we want to prove strong convexity on the slice $\sum_i \mu_i a_i = \sum_j \beta_j b_j$, we compute

$$\begin{aligned} (a,b)^T M (a,b) &= \mathbb{E}\left[D_{i,j}(a_i+b_j)^2\right] \\ &\geq e^{-B/\varepsilon}\mathbb{E}\left[(a_i+b_j)^2\right]. \end{aligned}$$

We add that

$$\mathbb{E}\left[(a_i+b_j)^2\right] = \sum_i \mu_i a_i^2 + \sum_j \beta_j b_j^2 + 2(\sum_i \mu_i a_i)(\sum_j \beta_j b_j)$$

thus

$$\mathbb{E}\left[(a_i+b_j)^2\right] = \sum_i (\mu_i + \mu_i^2) a_i^2 + \sum_j (\beta_j + \beta_j^2) b_j^2$$

since we are on the slice. So $M \succcurlyeq \lambda \mathrm{Id}$ and finally $-F$ is $\lambda-$strongly convex with

$$\lambda = \frac{\min_{i,j}\{\mu_i,\beta_j\}}{\varepsilon} e^{-B/\varepsilon}.$$

4. We compute the gradients of F:

$$\frac{\partial F}{\partial a_i}(a,b) = \mu_i - \mu_i \sum_{j=1}^{J} \beta_j D_{i,j}(a,b),$$
$$\frac{\partial F}{\partial b_j}(a,b) = \nu_j^*(-b/\eta) - \beta_j \sum_{i=1}^{I} \mu_i D_{i,j}(a,b),$$

with $D_{i,j}(a,b) = e^{\frac{a_i + b_j - C_{i,j}}{\varepsilon}}$. If we take i and j to be independent random variables following the laws (μ_i) and (β_j) respectively, we have the desired expression for the gradients.

□

Proof of Lemma 1. With the initial conditions, we guarantee that $0 \leq G_a^0 \leq 1$ and $0 \leq G_b^0 \leq f_j \leq e^m$. At each timestep t, we have

$$\|\nabla F_{i,j}^t\|^2 \leq \max\{2e^{2m}, 2(D_{i,j}^t)^2\},$$

with i, j being two independent random variables following the laws μ and β respectively. If $D_{i,j}^t \geq e^m$, then $G_a + G_b \leq 0$ and

$$D_{i,j}^{t+1} = D_{i,j}^t e^{\frac{G_a + G_b}{\varepsilon}} \leq D_{i,j}^t.$$

Moreover if $D_{i,j}^t \leq e^m$ then $\|\nabla F_{i,j}^t\|^2 \leq 1 + e^{2m}$ thus $\mathbb{E}\left[\max\{2e^{2m}, (D_{i,j}^t)^2\}\right]$ is a decreasing function of t. Thus we have the bound

$$\mathbb{E}\left[\|\nabla_a F_{i,j}(a^t, b^t)\|^2 + \|\nabla_b F_{i,j}(a^t, b^t)\|^2\right] \leq 2e^{2m}.$$

□

Proof of Lemma 2. We first assume that (a,b) and (a^*, b^*) are on the slice $\{\sum_i \mu_i a_i = \sum_j \beta_j b_j\}$. By strong convexity of $-F$ on this slice we have

$$|b - b^*|^2 \leq \frac{2(F(a^*, b^*) - F(a,b))}{\lambda}. \tag{5}$$

We remark that the function $g : b \mapsto KL(\nu(b^*), \nu(b))$ verifies

$$\begin{aligned}
\partial_i g(b) &= -\sum_j \nu_j(b^*) \partial_i \log \nu_j(b), \\
&= -\sum_j \nu_j(b^*) \nu_j(b)^{-1} \partial_i \nu_j(b), \\
&= \frac{1}{\eta} \sum_j \nu_j(b^*) \nu_j(b)^{-1} \nu_i(\delta_{ij} - \nu_j(b)), \\
&= \frac{\nu_i(b^*) - \nu_i(b)}{\eta - \varepsilon},
\end{aligned}$$

thus

$$\begin{aligned}
\partial_i \partial_j g(b) &= -\frac{\partial_j \nu_i(b)}{\eta - \varepsilon}, \\
&= -\frac{\nu_j(b)(\delta_{ij} - \nu_i(b))}{\eta - \varepsilon},
\end{aligned}$$

so the Hessian matrix $\nabla^2 g(b)$ of g is a sum of a diagonal matrix with the negative values $-\nu_j(b)/(\eta - \varepsilon)$ and the one-rank matrix $(\nu_j(b)\nu_i(b)/(\eta - \varepsilon))_{i,j}$. Hence the eigenvalues of $\nabla^2 g(b)$ are contained in $[-1/(\eta - \varepsilon), 1/(\eta - \varepsilon)]$, thus Taylor's inequality gives

$$\begin{aligned}
g(b) &\leq g(b^*) + |b - b^*| \|\nabla g(b^*)\| + \frac{|b - b^*|^2}{2(\eta - \varepsilon)}, \\
&\leq \frac{|b - b^*|^2}{2(\eta - \varepsilon)},
\end{aligned}$$

because $g(b^*) = 0$ and $\nabla g(b^*) = 0$. We complete the proof with (5). For the case where the vector (a,b) or (a^*, b^*) is not on the slice $\{\sum_i \mu_i a_i = \sum_j \beta_j b_j\}$, we note that adding a constant vector $c = (c_1, \ldots, c_1)$ to b does not change the value of $\nu(b)$, and that F is invariant by translation in the direction $(-c, +c)$. With $c_1 = \left(\sum_i \mu_i a_i - \sum_j \beta_j b_j\right)/2$, the vectors $(a', b') = (a + c, b - c)$ are on the slice and verify $\nu(b') = \nu(b)$ and $F(a', b') = F(a, b)$. Hence the result for (a', b') implies the result for (a, b). □

References

Rigollet, P. and Weed, J. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11-12):1228–1235, 2018.