Supplementary Material for **Adaptive Metropolis with Online Relabeling**

Notations and assumptions

By convention, vectors $x \in \mathbb{R}^d$ are column vectors. x^T is the transpose vector of x. We fix a norm $\|\cdot\|$ on vectors and will also denote by $\|\cdot\|$ the derived norm for matrices.

Let π be a probability density with respect to (w.r.t.) the Lebesgue measure on $\mathcal{X}$ where $\mathcal{X} \subseteq \mathbb{R}^d$ is measurable ($\mathbb{R}^d$ is endowed with its Borel σ-field). It is assumed that

(a) π in invariant under the action of $\mathcal{P}$, a finite group of $d \times d$ block permutation matrices.

(b) π has finite second moment.

Let $\mathcal{C}_d^+$ be the set of the real $d \times d$ (symmetric) positive definite matrices. For any $\theta = (\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{C}_d^+$ and $x \in \mathcal{X}$, define the quadratic loss

$$L_\theta(x) = (x-\mu)^T \Sigma^{-1} (x-\mu) \tag{1}$$

and the set

$$V_\theta = \{x \in \mathcal{X} : L_\theta(x) = \min_{P \in \mathcal{P}} L_\theta(Px)\}.$$

Observe that for any $\theta \in \mathbb{R}^d \times \mathcal{C}_d^+$, V_θ is measurable.

For any $\theta \in \mathbb{R}^d \times \mathcal{C}_d^+$, let π_θ be the probability density on $\mathbb{R}^d$, defined by

$$\pi_\theta(x) = Z_\theta^{-1} \mathbb{1}_{V_\theta}(x) \pi(x), \quad \text{where} \quad Z_\theta = \int_{V_\theta} \pi(x) dx.$$

Under the assumptions on π, π_θ has an expectation

$$\mu_{\pi_\theta} = \int x \, \pi_\theta(x) dx$$

and a covariance matrix

$$\Sigma_{\pi_\theta} = \int (x - \mu_{\pi_\theta})(x - \mu_{\pi_\theta})^T \pi_\theta(x) dx.$$

Define the function $w : \mathbb{R}^d \times \mathcal{C}_d^+ \to \mathbb{R}$ by

$$w(\theta) = -\int \log \mathcal{N}(x|\theta) \, \pi_\theta(x) dx.$$

Finally, denote by $\mathcal{M}_d$ the set of $d \times d$ real matrices. Define the function $h : \mathbb{R}^d \times \mathcal{C}_d^+ \to \mathbb{R}^d \times \mathcal{M}_d$ by

$$h(\theta) = \left((\mu_{\pi_\theta} - \mu), \Sigma_{\pi_\theta} - \Sigma + (\mu_{\pi_\theta} - \mu)(\mu_{\pi_\theta} - \mu)^T \right). \tag{2}$$

The set $\mathbb{R}^d \times \mathcal{M}_d$ is endowed with the scalar product given by

$$x = (\mu_1, M_1), y = (\mu_2, M_2) \in \mathbb{R}^d \times \mathcal{M}_d, \qquad \langle x, y \rangle = \mu_1^T \mu_2 + \text{Trace}\left(M_1^T M_2\right).$$

1 Main result

We first prove that w is positive on the set Θ defined by

$$\Theta = \{\theta = (\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{C}_d^+ : \forall P \in \mathcal{P}, P^T \Sigma P \neq \Sigma \text{ or } \mu \neq P\mu\}, \tag{3}$$

and for any $\theta \in \Theta$, $w(\theta)$ is, up to a constant, the Kullback-Leibler divergence between π_θ and the Gaussian distribution $\mathcal{N}(\cdot|\theta)$.

Proposition 1. *For any $\theta \in \Theta$,*

$$w(\theta) = \int \log \frac{\pi_\theta(x)}{\mathcal{N}(x|\theta)}\, \pi_\theta(x) dx - \left(\log |\mathcal{P}| + \int \log \pi(x) \pi(x) dx \right),$$

where $|\mathcal{P}|$ denotes the cardinal of $\mathcal{P}$.

Proposition 2 shows that for any $\theta \in \Theta$, w is similar to a distortion measure in vector quantization [1].

Proposition 2. *For any $\theta \in \Theta$,*

$$w(\theta) = \frac{1}{2} \ln \det(\Sigma) + \frac{1}{2} \int \min_{P \in \mathcal{P}} L_{(P\mu, P\Sigma P^T)}(x)\, \pi(x) dx.$$

Finally, Proposition 3 and Corollary 1 show that on Θ, w is a natural Lyapunov function for the mean field h given by (2).

Proposition 3. *The function w is continuously differentiable on Θ and for any $\theta \in \Theta$,*

$$\begin{aligned} \nabla_\mu w(\theta) &= -\Sigma^{-1}(\mu_{\pi_\theta} - \mu), \\ \nabla_\Sigma w(\theta) &= -\frac{1}{2} \Sigma^{-1} \left(\Sigma_{\pi_\theta} - \Sigma + (\mu_{\pi_\theta} - \mu)(\mu_{\pi_\theta} - \mu)^T \right) \Sigma^{-1}. \end{aligned}$$

Corollary 1. *For any $\theta \in \Theta$,*

$$\langle \nabla w(\theta), h(\theta) \rangle \leq 0,$$

and $\langle \nabla w(\theta), h(\theta) \rangle = 0$ iff $\mu = \mu_{\pi_\theta}$ and $\Sigma = \Sigma_{\pi_\theta}$.

Corollary 1 is equivalent to Proposition 2 in the main paper.

Proofs

1.1 Proof of Proposition 1

We start by proving a lemma. Let

$$PV_\theta = \{Px : x \in V_\theta\}. \tag{4}$$

Lemma 1. *For any $\theta \in \Theta$, the sets $\{PV_\theta, P \in \mathcal{P}\}$ cover $\mathcal{X}$ and for any $P, Q \in \mathcal{P}$, $P \neq Q$, the Lebesgue measure of $PV_\theta \cap QV_\theta$ is zero.*

Therefore, $Z_\theta = |\mathcal{P}|^{-1}$ for any $\theta \in \Theta$.

Proof. Let $\theta \in \Theta$. We first prove that for any $P, Q \in \mathcal{P}$ and $P \neq Q$, the Lebesgue measure of $PV_\theta \cap QV_\theta$ is zero. Observe that $PV_\theta \cap QV_\theta \subseteq \{x : L_\theta(P^T x) = L_\theta(Q^T x)\}$ and $L_\theta(P^T x) = L_\theta(Q^T x)$ iff

$$(x - P\mu)^T P \Sigma^{-1} P^T (x - P\mu) = (x - Q\mu)^T Q \Sigma^{-1} Q^T (x - Q\mu),$$

or, equivalently,

$$x^T \left(P\Sigma^{-1}P^T - Q\Sigma^{-1}Q^T \right) x - 2\mu^T \left(\Sigma^{-1}P^T - \Sigma^{-1}Q^T \right) x = 0.$$

Then $\{x : L_\theta(P^T x) = L_\theta(Q^T x)\}$ is either a quadratic or a linear surface, and thus of Lebesgue measure zero, except if both $\Sigma^{-1} = R^T \Sigma^{-1} R$ and $\mu = R\mu$ with $R = Q^T P$. Since $\mathcal{P}$ is a group, $R \in \mathcal{P}$ and the definition of Θ now guarantees that these two conditions never simultaneously hold when $\theta \in \Theta$.

We now prove that $\mathcal{X} \subseteq \bigcup_{P\in\mathcal{P}} PV_\theta$. For any $x \in \mathcal{X}$, there exists $P \in \mathcal{P}$ such that $L_\theta(Px) = \min_{Q\in\mathcal{P}} L_\theta(Qx)$. Then, $x \in P^T V_\theta$ and this concludes the proof since $\mathcal{P}$ is a group.

Let $P \in \mathcal{P}$. Observe that since π is invariant under the action of $\mathcal{P}$,

$$\int_{V_\theta} \pi(y)dy = \int_{V_\theta} \pi(Py)dy = \int_{PV_\theta} \pi(x)dx.$$

Then, since $\mathrm{Leb}(PV_\theta \cap QV_\theta) = 0$ for any $P \neq Q$ and $\mathcal{X} = \bigcup_{P\in\mathcal{P}} PV_\theta$,

$$Z_\theta = \int_{V_\theta} \pi(y)dy = \frac{1}{|\mathcal{P}|} \sum_{P\in\mathcal{P}} \int_{PV_\theta} \pi(x)dx = \frac{1}{|\mathcal{P}|} \int \pi(x)dx = \frac{1}{|\mathcal{P}|}.$$

□

Proof. (of Proposition 1) Since $\pi(Px) = \pi(x)$ for any $x \in \mathcal{X}$ and $P \in \mathcal{P}$,

$$\int_{V_\theta} \log \pi(y)\, \pi(y)dy = \int_{V_\theta} \log \pi(Py)\, \pi(Py)dy = \int_{PV_\theta} \log \pi(x)\, \pi(x)dx.$$

Then, by Lemma 1, for any $\theta \in \Theta$,

$$\int_{V_\theta} \log \pi(y)\, \pi(y)dy = \frac{1}{|\mathcal{P}|} \sum_{P\in\mathcal{P}} \int_{PV_\theta} \log \pi(x)\, \pi(x)dx = \frac{1}{|\mathcal{P}|} \int \log \pi(x)\pi(x)dx.$$

Since $Z_\theta = 1/|\mathcal{P}|$ by Lemma 1, this implies that

$$-\int \log \pi_\theta(x)\, \pi_\theta(x)dx = -\log|\mathcal{P}| - |\mathcal{P}| \int_{V_\theta} \log \pi(x)\, \pi(x)dx = -\log|\mathcal{P}| - \int \log \pi(x)\pi(x)dx,$$

thus showing that for any $\theta \in \Theta$,

$$\int \log \frac{\pi_\theta(x)}{\pi(x)}\, \pi_\theta(x)dx = \log|\mathcal{P}|,$$

and

$$\int \log \frac{\pi_\theta(x)}{\mathcal{N}(x|\theta)}\, \pi_\theta(x)dx = w(\theta) + \log|\mathcal{P}| + \int \log \pi(x)\pi(x)dx.$$

□

1.2 Proof of Proposition 2 (Proposition 3 in the main paper)

Let $\theta \in \Theta$. By definition of w and by Lemma 1,

$$w(\theta) = \frac{1}{2}\ln\det(\Sigma) + \frac{|\mathcal{P}|}{2} \int_{V_\theta} L_\theta(x)\pi(x)dx$$

where V_θ and L_θ are given resp. by (4) and (1) and $|\mathcal{P}|$ denotes the cardinal of $\mathcal{P}$. We have

$$|\mathcal{P}| \int_{V_\theta} L_\theta(x)\pi(x)dx = \sum_{P\in\mathcal{P}} \int_{V_\theta} L_\theta(x)\pi(x)dx = \sum_{P\in\mathcal{P}} \int_{PV_\theta} L_\theta(P^T x)\pi(x)dx,$$

where we use that π is invariant under the action of $\mathcal{P}$. In addition, by definition,

$$PV_\theta = \{x \in \mathcal{X} : L_\theta(P^T x) = \min_{Q\in\mathcal{P}} L_\theta(Qx)\}.$$

Then by using Lemma 1,

$$|\mathcal{P}| \int_{V_\theta} L_\theta(x)\pi(x)dx = \sum_{P\in\mathcal{P}} \int_{PV_\theta} \min_{Q\in\mathcal{P}} L_\theta(Qx)\pi(x)dx = \int \min_{Q\in\mathcal{P}} L_\theta(Qx)\pi(x)dx.$$

Finally, by definition of L_θ, $L_\theta(Qx) = L_{(Q^T\mu, Q^T\Sigma Q)}(x)$, and this concludes the proof.

1.3 Proof of Proposition 3

We start by two lemmas. Lemma 2 is established for generic loss functions L_θ and a generic open set Θ. Its proof is adapted from [1, Lemma 4.10, page 44]. We then show in Lemma 3 that this result applies to the loss function given by (1) and the set Θ given by (3).

Lemma 2. *Let Θ be an open subset of $\mathbb{R}^\ell$, r be a positive integer and $\mathcal{O} \subseteq \Theta^r$ be an open set. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a measurable set and π be a probability density w.r.t. the Lebesgue measure on $\mathcal{X}$. Let $\{L_\theta, \theta \in \Theta\}$ be a family of loss functions $L_\theta : \mathcal{X} \to \mathbb{R}$, satisfying*

1. *For π-almost every x, $\theta \mapsto L_\theta(x)$ is C^1 on Θ and for any $\theta \in \Theta$, there exists $h_0 > 0$ such that*
$$\int \sup_{\|h\| \le h_0} \frac{1}{\|h\|} |h^T \nabla_\theta L_\theta(x)| \, \pi(x) dx < \infty.$$
2. *For any $\theta \in \Theta$, there exists $h_0 > 0$ such that*
$$\int \sup_{\|h\| \le h_0} \frac{|L_{\theta+h}(x) - L_\theta(x)|}{\|h\|} \, \pi(x) dx < \infty.$$
3. *For any $\boldsymbol{\theta} = (\theta_1, ..., \theta_r) \in \mathcal{O}$, the sets*
$$V_{\theta_i} = \{x \in \mathcal{X} : L_{\theta_i}(x) \le \min_j L_{\theta_j}(x)\}$$
are measurable, cover $\mathcal{X}$ and for any $i \ne j$, the Lebesgue measure of $V_{\theta_i} \cap V_{\theta_j}$ is zero.

For $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_r) \in \mathcal{O}$ define the function $\psi : \Theta^r \to \mathbb{R}$
$$\psi(\boldsymbol{\theta}) = \int \min_{1 \le i \le r} L_{\theta_i}(x) \, \pi(x) dx.$$
Then ψ is differentiable on $\mathcal{O}$ and for $1 \le i \le r$,
$$\nabla_{\theta_i} \psi(\boldsymbol{\theta}) = \int_{V_{\theta_i}} \nabla_{\theta_i} L_{\theta_i}(x) \, \pi(x) dx.$$

Proof. (of Lemma 2) Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_r) \in \mathcal{O}$. Set
$$d(x, \boldsymbol{\theta}) = \min_{1 \le i \le r} L_{\theta_i}(x).$$
By definition of the function ψ
$$\psi(\boldsymbol{\theta} + \mathbf{h}) - \psi(\boldsymbol{\theta}) = \int \big(d(x, \boldsymbol{\theta} + \mathbf{h}) - d(x, \boldsymbol{\theta}) \big) \, \pi(x) dx. \tag{5}$$
We prove that $\lim_{\|h\| \to 0} \|h\|^{-1} \left(\psi(\boldsymbol{\theta} + \mathbf{h}) - \psi(\boldsymbol{\theta}) - \sum_{i=1}^r \int_{V_{\theta_i}} \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \, \pi(x) dx \right) = 0$ by applying the dominated convergence theorem.

By Assumption 3,
$$\begin{aligned} \psi(\boldsymbol{\theta} + \mathbf{h}) - \psi(\boldsymbol{\theta}) \quad &- \quad \sum_{i=1}^r \int_{V_{\theta_i}} \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \, \pi(x) dx \\ &= \quad \sum_{i=1}^r \int_{V_{\theta_i}} \big(d(x, \boldsymbol{\theta} + \mathbf{h}) - d(x, \boldsymbol{\theta}) - \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \big) \, \pi(x) dx. \end{aligned}$$
Set
$$V^\circ_{\theta_i} = \{x \in \mathcal{X} : L_{\theta_i}(x) < \min_{j \ne i} L_{\theta_j}(x)\}$$

and note that $V_{\theta_i} \setminus V^\circ_{\theta_i}$ has measure zero under Assumption 3. Then

$$\psi(\boldsymbol{\theta}+\mathbf{h}) - \psi(\boldsymbol{\theta}) - \sum_{i=1}^r \int_{V_{\theta_i}} \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \, \pi(x) dx$$
$$= \sum_{i=1}^r \int_{V^\circ_{\theta_i}} \big(d(x, \boldsymbol{\theta}+\mathbf{h}) - d(x, \boldsymbol{\theta}) - \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \big) \, \pi(x) dx.$$

Let $x \in V^\circ_{\theta_i}$; under Assumption 1, $\theta \mapsto L_\theta(x)$ is continuous on Θ and there exists ε_x such that

$$\|h\| \leq \varepsilon_x \Rightarrow d(x, \boldsymbol{\theta}+\mathbf{h}) = L_{\theta_i + h_i}(x).$$

Then, by Assumption 1,

$$d(x, \boldsymbol{\theta}+\mathbf{h}) - d(x, \boldsymbol{\theta}) - \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle = L_{\theta_i+h_i}(x) - L_{\theta_i}(x) - \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle = C(\theta_i, x, h_i)$$

with $\|h_i\|^{-1} C(\theta_i, x, h_i) \to 0$ when $\|h_i\| \to 0$. Hence, we proved that for any $i \leq r$ and any $x \in V^\circ_{\theta_i}$,

$$\lim_{\|h\| \to 0} \|h\|^{-1} \big(d(x, \boldsymbol{\theta}+\mathbf{h}) - d(x, \boldsymbol{\theta}) - \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \big) = 0.$$

We now prove that there exists h_0 such that

$$\int \sup_{\|h\| \leq h_0} \|h\|^{-1} \Big| d(x, \boldsymbol{\theta}+\mathbf{h}) - d(x, \boldsymbol{\theta}) - \sum_{i=1}^r \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \, \mathbb{1}_{V_{\theta_i}}(x) \Big| \, \pi(x) dx < +\infty. \tag{6}$$

First remark that for all $z, \mathbf{a} = (a_1, \cdots, a_r)$, $\mathbf{b} = (b_1, \cdots, b_r)$,

$$|d(z, \mathbf{a}+\mathbf{b}) - d(z, \mathbf{a})| \leq \max_{1 \leq i \leq r} |L_{a_i+b_i}(z) - L_{a_i}(z)|. \tag{7}$$

Indeed, assume without loss of generality that $d(z, \mathbf{a}) \leq d(z, \mathbf{a}+\mathbf{b})$ and let i be such that $d(z, \mathbf{a}) = L_{a_i}(z)$, then by definition of the distance d, $d(z, \mathbf{a}+\mathbf{b}) \leq L_{a_i+b_i}(z)$, which proves Eq. (7). Now, the proof of (6) is a consequence of Assumptions 1 and 2 and the inequality

$$\max_{1 \leq i \leq r} |L_{a_i+b_i}(z) - L_{a_i}(z)| \leq \sum_{i=1}^r |L_{a_i+b_i}(z) - L_{a_i}(z)|.$$

□

Lemma 3. *The quadratic loss function given by (1), the set Θ given by (3) and the open set*

$$\mathcal{O} = \{(P\mu, P\Sigma P^T) : P \in \mathcal{P}, (\mu, \Sigma) \in \Theta\}$$

satisfy the assumptions of Lemma 2.

Proof. (of Lemma 3) When taking derivatives with respect to a matrix, we shall use the "vec" notation during computations. For a $d \times d$ matrix A, its vectorized form $\text{vec}(A)$ is a d^2 vector such that $\text{vec}(A)$ stacks the columns of A on top of one another. In general, we refer to [2] for matrix algebra notions.

We check the conditions of Lemma 2. Denote by r the cardinality of $\mathcal{P}$ and set $\mathcal{P} = (I_d, P_2, \cdots, P_r)$. We set

$$\mathcal{O} = \{(\theta_1, \cdots, \theta_r) \in \Theta^r : \theta_i = (P_i\mu, P_i\Sigma P_i^T), \forall i \geq 1\}.$$

Note that $L_{\theta_i}(x) = L_{\theta_1}(P_i^T x)$ and $V_{\theta_i} = P_i V_{\theta_1}$.

We have

$$(\mu, \Sigma) \mapsto (x-\mu)^T \Sigma^{-1} (x-\mu) = \frac{1}{\det \Sigma} (x-\mu)^T \text{Adjugate}(\Sigma)(x- \; u)$$

so that $\theta \mapsto L_\theta(x)$ is a rational function in the coefficients of μ and Σ whose denominator $\det \Sigma > 0$. In addition,

$$\sup_{\|h\|\le h_0} \frac{1}{\|h\|} \left| h^T \nabla_\theta L_\theta(x) \right| \le \|\nabla_\theta L_\theta(x)\| \le \|\nabla_\mu L_\theta(x)\| + \|\nabla_\Sigma L_\theta(x)\|.$$

The RHS is at most quadratic in x (for fixed θ). Under the stated assumptions on π, the RHS is π-integrable. This proves Assumption 1.

We now prove Assumption 2. Let $\theta \in \Theta$ and set $\Delta\theta = (\Delta\mu, \Delta\Sigma)$. By standard algebra, we have

$$(\Sigma + \Delta\Sigma)^{-1} = \Sigma^{-1} - \Sigma^{-1}\,\Delta\Sigma\,\Sigma^{-1} + o(\|\Delta\Sigma\|)$$

for any matrix $\Delta\Sigma$ such that $\Sigma + \Delta\Sigma$ is invertible. Therefore,

$$L_{\theta+\Delta\theta}(x) - L_\theta(x) = -2(\Delta\mu)^T\Sigma^{-1}(x-\mu) - (x-\mu)^T\Sigma^{-1}\,\Delta\Sigma\,\Sigma^{-1}(x-\mu) + \Xi(x,\theta,\Delta\theta),$$

for some function $\Xi(x,\theta,\Delta\theta)$ such that

$$|\Xi(x,\theta,\Delta\theta)| \le C(\theta)\|x\|^2\|\Delta\theta\|^2$$

and some constant $C(\theta)$ (depending upon θ but independent of x and $\Delta\theta$). The proof is concluded since $\int \|x\|^2\pi(x)dx < +\infty$.

Finally, the sets V_{θ_i} are measurable for any $\theta_1, \cdots, \theta_r \in \Theta$ since $(x,\theta) \mapsto L_\theta(x)$ is continuous on $\mathcal{X} \times \Theta$. The proof of Assumption 3 is then concluded by application of Lemma 1. □

We finally turn to proving Proposition 3.

Proof. (of Proposition 3) Let r denote the cardinality of $\mathcal{P}$ and set $\mathcal{P} = (I_d, P_2, \cdots, P_r)$. Let $\theta \in \Theta$. By Proposition 2, we have

$$w(\theta) = \frac{1}{2}\ln\det(\Sigma) + \frac{1}{2}\int \min_{1\le i\le r} L_{\theta_i}(x)\,\pi(x)dx,$$

where $\theta_i = (P_i\mu, P_i\Sigma^{-1}P_i^T)$.

We first consider the derivative w.r.t. μ. We have

$$\nabla_\mu w(\theta) = \frac{1}{2}\nabla_\mu \int \min_{1\le i\le r} L_{\theta_i}(x)\,\pi(x)dx.$$

By Lemmas 2 and 3 and the chain rule, we have

$$\begin{aligned}\nabla_\mu w(\theta) &= \frac{1}{2}\sum_{i=1}^r P_i^T \int_{\{x: L_{\theta_i}(x)\le \min_j L_{\theta_j}(x)\}} \nabla_{\mu_i}\left[(x-\mu_i)P_i\Sigma^{-1}P_i^T(x-\mu_i)\right]_{\mu_i = P_i\mu}\,\pi(x)dx \\ &= -\Sigma^{-1}\sum_{i=1}^r \int_{\{x: L_{\theta_i}(x)\le \min_j L_{\theta_j}(x)\}} (P_i^T x - \mu)\,\pi(x)dx\end{aligned}$$

By definition of P_iV_θ (see (4)),

$$\{x : L_{\theta_i}(x) \le \min_j L_{\theta_j}(x)\} = P_iV_\theta.$$

Hence, by Lemma 1 and since π is invariant under action of $\mathcal{P}$, we have

$$\nabla_\mu w(\theta) = -\Sigma^{-1}\sum_{i=1}^r \int_{V_\theta}(x-\mu)\,\pi(x)dx = -\Sigma^{-1}\int (x-\mu)[r\pi(x)\mathbb{1}_{V_\theta}(x)]dx = -\Sigma^{-1}\left(\mu_{\pi_\theta} - \mu\right),$$

where we used the definition of μ_{π_θ}.

We now consider the derivative w.r.t. Σ, that we will derive in a similar manner. We refer to [2] for matrix algebra notions such as Kronecker products. First remark that, by standard algebra and since Σ is symmetric,

$$\nabla_{\mathrm{vec}(\Sigma)}\ln\det\Sigma = \mathrm{vec}(\Sigma^{-1}).$$

Then recall that

$$\nabla_{\text{vec}(\Sigma)}(x-\mu)\Sigma^{-1}(x-\mu) = -\Sigma^{-1}(x-\mu)\otimes\Sigma^{-1}(x-\mu).$$

Now, using Lemmas 2 and 3 along with the chain rule, we have

$$\begin{aligned}
\nabla_{\text{vec}(\Sigma)}w(\theta) - \frac{1}{2}\text{vec}(\Sigma^{-1}) &= \frac{1}{2}\sum_{i=1}^{r}(P_i\otimes P_i)^T\int_{P_iV_\theta}\nabla_{\text{vec}(\Sigma_i)}\left[(x-P_i\mu)^T\Sigma_i^{-1}(x-P_i\mu)\right]_{\Sigma_i=P_i\Sigma P_i^T}\pi(x)dx \\
&= -\frac{1}{2}\sum_{i=1}^{r}(P_i^T\otimes P_i^T)\int_{P_iV_\theta}\left[P_i\Sigma^{-1}P_i^T(x-P_i\mu)\right]\otimes\left[P_i\Sigma^{-1}P_i^T(x-P_i\mu)\right]\pi(x)dx \\
&= -\frac{1}{2}\sum_{i=1}^{r}\int_{P_iV_\theta}\left[\Sigma^{-1}(P_i^Tx-\mu)\right]\otimes\left[\Sigma^{-1}(P_i^Tx-\mu)\right]\pi(x)dx \\
&= -\frac{1}{2}(\Sigma^{-1}\otimes\Sigma^{-1})\sum_{i=1}^{r}\int_{P_iV_\theta}[P_i^Tx-\mu]\otimes[P_i^Tx-\mu]\pi(x)dx
\end{aligned}$$

where we used the identities $(A\otimes B)^T = A^T\otimes B^T$ and $(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$. A change of variables now leads to

$$\begin{aligned}
\nabla_{\text{vec}(\Sigma)}w(\theta) - \frac{1}{2}\text{vec}(\Sigma^{-1}) &= -\frac{1}{2}(\Sigma^{-1}\otimes\Sigma^{-1})\sum_{i=1}^{r}\int_{V_\theta}(x-\mu)\otimes(x-\mu)\pi(x)dx \\
&= -\frac{1}{2}(\Sigma^{-1}\otimes\Sigma^{-1})\int(x-\mu_{\pi_\theta}+\mu_{\pi_\theta}-\mu)\otimes(x-\mu_{\pi_\theta}+\mu_{\pi_\theta}-\mu)[r\pi(x)\mathbb{1}_{V_\theta}(x)]dx \\
&= -\frac{1}{2}(\Sigma^{-1}\otimes\Sigma^{-1})\left(\int(x-\mu_{\pi_\theta})\otimes(x-\mu_{\pi_\theta})\pi_\theta(x)dx + (\mu_{\pi_\theta}-\mu)\otimes(\mu_{\pi_\theta}-\mu)\right) \\
&= -\frac{1}{2}(\Sigma^{-1}\otimes\Sigma^{-1})\text{vec}(\Sigma_{\pi_\theta}+(\mu_{\pi_\theta}-\mu)(\mu_{\pi_\theta}-\mu)^T)
\end{aligned}$$

where we used the distributivity of the Kronecker product, Lemma 1 and the definitions of μ_{π_θ} and Σ_{π_θ}. Finally, the identity $\text{vec}(AXB) = (B^T\otimes A)\text{vec}(X)$ allows us to write

$$\nabla_{\text{vec}(\Sigma)}w(\theta) = -\frac{1}{2}\text{vec}\big(\Sigma^{-1}[\Sigma_{\pi_\theta}-\Sigma+(\mu_{\pi_\theta}-\mu)(\mu_{\pi_\theta}-\mu)^T]\Sigma^{-1}\big).$$

□

1.4 Proof of Corollary 1 (Proposition 1 in the main paper)

Let $\theta\in\mathbb{R}^d\times\mathcal{C}_d^+$. By definition of the scalar product on $\mathbb{R}^d\times\mathcal{M}_d$ we have

$$\begin{aligned}
\langle\nabla w(\theta), h(\theta)\rangle = &-(\mu_{\pi_\theta}-\mu)^T\,\Sigma^{-1}\,(\mu_{\pi_\theta}-\mu)^T \\
&-\frac{1}{2}\text{Trace}\left(\Sigma^{-1}[\Sigma_{\pi_\theta}-\Sigma+(\mu_{\pi_\theta}-\mu)(\mu_{\pi_\theta}-\mu)^T]\Sigma^{-1}[\Sigma_{\pi_\theta}-\Sigma+(\mu_{\pi_\theta}-\mu)(\mu_{\pi_\theta}-\mu)^T]\right).
\end{aligned}$$

The first term of the right-hand side is negative since $\Sigma^{-1}\in\mathcal{C}_d^+$, and this term is null iff $\mu=\mu_{\pi_\theta}$. For the second term, note that since $(A,B)\mapsto\text{Trace}(A^TB)$ is a scalar product on $\mathcal{M}_d$, $\text{Trace}A^TA\geq 0$. This yields

$$\begin{aligned}
&\text{Trace}\left(\Sigma^{-1}[\Sigma_{\pi_\theta}-\Sigma+(\mu_{\pi_\theta}-\mu)(\mu_{\pi_\theta}-\mu)^T]\Sigma^{-1}[\Sigma_{\pi_\theta}-\Sigma+(\mu_{\pi_\theta}-\mu)(\mu_{\pi_\theta}-\mu)^T]\right) \\
&\quad= \text{Trace}\left(\Sigma^{-1/2}[\Sigma_{\pi_\theta}-\Sigma+(\mu_{\pi_\theta}-\mu)(\mu_{\pi_\theta}-\mu)^T]\Sigma^{-1}[\Sigma_{\pi_\theta}-\Sigma+(\mu_{\pi_\theta}-\mu)(\mu_{\pi_\theta}-\mu)^T]\Sigma^{-1/2}\right)\geq 0,
\end{aligned}$$

and when $\mu=\mu_{\pi_\theta}$, this term is null iff $\Sigma=\Sigma_{\pi_\theta}$.

References

[1] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer-Verlag, 2000.

[2] J.W. Brewer. Kronecker products and matrix calculus in system theory. *IEEE Transactions on Circuits and Systems*, 25:772–781, 1978.