Robust Bayesian Recourse (Supplementary Material)

Tuan-Duy H. Nguyen[1] **Ngoc Bui**[1] **Duy Nguyen**[1] **Man-Chung Yue**[2] **Viet Anh Nguyen**[1]

[1]VinAI Research, Vietnam
[2]The University of Hong Kong

A PROOFS OF SECTION 5

Lemma 5.3 (re-stated). *There exists a distribution $\mathbb{Q}_0^\star$ that solves* (3) *and is a mixture of at most N_0 Gaussian components. Moreover, problem* (3) *is equivalent to a separable problem of the form*

$$\max\left\{L(x,\mathbb{Q}_0):\mathbb{Q}_0\in\mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}_0^\sigma)\right\}=\left\{\begin{array}{ll}\max & \frac{1}{N_0}\sum_{i\in\mathcal{I}_0} f(x|\mu_i,\Sigma_i)\\ \text{s.t.} & (\mu_i,\Sigma_i)\in\mathbb{R}^p\times\mathbb{S}_{\geq\sigma}^p\\ & c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2\bar{I}))\leq\varepsilon_0\quad\forall i\in\mathcal{I}_0.\end{array}\right.$$

An analogous result holds for problem (4) *with the corresponding subscript $y=1$.*

Proof of Lemma 5.3. There exists a distribution $\mathbb{Q}_0^\star$ that solves (3) and is a mixture of at most N_0 Gaussian components. Moreover, problem (3) is equivalent to a separable problem of the form

$$\begin{array}{l}\max\left\{L(x,\mathbb{Q}_0):\mathbb{Q}_0\in\mathbb{B}_{\varepsilon_0}(\widehat{\mathbb{P}}_0^\sigma)\right\}\\ =\left\{\begin{array}{ll}\max & \frac{1}{N_0}\sum_{i\in\mathcal{I}_0} f(x|\mu_i,\Sigma_i)\\ \text{s.t.} & (\mu_i,\Sigma_i)\in\mathbb{R}^p\times\mathbb{S}_{\geq\sigma}^p\\ & c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2\bar{I}))\leq\varepsilon_0\quad\forall i\in\mathcal{I}_0.\end{array}\right.\end{array}$$

We use $\forall i$ implies $\forall i\in\mathcal{I}_0$, and $\sum_i$ is also taken over the same set. Given any x, the likelihood of x under any Gaussian mixture $\mathbb{Q}_0$ can be written using the corresponding measure ν_0 as

$$L(x,\mathbb{Q}_0)=\int_{\mathbb{R}^p\times\mathbb{S}_+^p} f(x|\mu,\Sigma)\nu_0(\mathrm{d}\mu,\mathrm{d}\Sigma).$$

Recall that $\Xi=\mathbb{R}^p\times\mathbb{S}_{\geq\sigma}^p$. Using the definition of the type-∞ Wasserstein, we find

$$\begin{array}{ll}& \mathbb{W}_c(\nu_0,\widehat{\nu}_0)\leq\varepsilon_0\\ \Leftrightarrow & \exists\lambda\in\Lambda(\nu_0,\widehat{\nu}_0)\text{ such that}\\ & \operatorname*{ess\,sup}_\lambda\left\{c((\mu,\Sigma),(\mu',\Sigma')):(\mu,\Sigma,\mu',\Sigma')\in\Xi\times\Xi\right\}\leq\varepsilon_0\\ \Leftrightarrow & \forall i\ \exists\lambda_i\in\mathcal{P}(\Xi)\text{ such that}\\ & \operatorname*{ess\,sup}_{\lambda_i}\left\{c((\mu,\Sigma),(\widehat{x}_i,\sigma I)):(\mu,\Sigma)\in\Xi\right\}\leq\varepsilon_0\\ \Leftrightarrow & \forall i\ \exists\lambda_i\in\mathcal{P}(\Xi)\text{ such that}\\ & c((\mu,\Sigma),(\widehat{x}_i,\sigma I))\leq\varepsilon_0\quad(\mu,\Sigma)\in\operatorname{supp}(\lambda_i),\end{array}$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

where the second equivalence follows from that $\widehat{\nu}_0 = \frac{1}{N_0}\sum_i \delta_{(\widehat{x}_i,\sigma^2 I)}$ and hence any $\lambda \in \Lambda(\nu_0, \widehat{\nu}_0)$ takes the form $\frac{1}{N_0}\sum_i \lambda_i \otimes \delta_{(\widehat{x}_i,\sigma^2 I)}$ for some probability measures $\lambda_i \in \mathcal{P}(\Xi)$, and the third equivalence follows from Lemma A.1. Hence, problem (3) is equivalent to

$$
\left\{\begin{array}{ll} \max & \int_{\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}} f(x|\mu,\Sigma)\nu_0(\mathrm{d}\mu,\mathrm{d}\Sigma) \\ \text{s.t.} & \nu_0 \in \mathcal{P}(\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}) \\ & \mathbb{W}_c(\nu_0, \widehat{\nu}_0) \leq \varepsilon_0 \end{array}\right.
$$

$$
= \left\{\begin{array}{lll} \max & \frac{1}{N_0}\sum_i \int_{\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}} f(x|\mu_i,\Sigma_i)\lambda_i(\mathrm{d}\mu_i,\mathrm{d}\Sigma_i) & \\ \text{s.t.} & \lambda_i \in \mathcal{P}(\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}) & \forall i \\ & c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2 I)) \leq \varepsilon_0 \quad \forall (\mu_i,\Sigma_i) \in \mathrm{supp}(\lambda_i) & \forall i. \end{array}\right.
$$

It is easy now to employ a greedy argument to show that the optimal solution for λ_i should be a Dirac delta distribution supported on one point in the space of $\mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma}$. This leads to the conclusion regarding the maximization problem (3).

An similar argument can be applied for the minimization problem (4), the detailed proof is omitted. □

Lemma A.1. *For any $\lambda \in \mathcal{P}(\Xi)$, $\widehat{x} \in \mathbb{R}^p$, $\sigma, \varepsilon > 0$ and any function $c : \Xi \times \Xi \to \mathbb{R}$ such that the map $(\mu,\Sigma) \mapsto c((\mu,\Sigma),(\widehat{x},\sigma^2 I))$ is continuous, we have $\operatorname{ess\,sup}_\lambda c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) \leq \varepsilon$ if and only if $c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) \leq \varepsilon$ for any $(\mu,\Sigma) \in \mathrm{supp}(\lambda)$.*

Proof of Lemma A.1. We first prove the "only if" direction. Suppose that there exists $(\mu',\Sigma') \in \mathrm{supp}(\lambda)$ such that

$$
c((\mu',\Sigma'),(\widehat{x},\sigma^2 I)) > \varepsilon.
$$

By continuity of the map $(\mu,\Sigma) \mapsto c((\mu,\Sigma),(\widehat{x},\sigma^2 I))$, there exists an open neighbourhood $U \subseteq \Xi$ containing (μ',Σ') such that

$$
c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) > \varepsilon \quad \forall (\mu,\Sigma) \in U.
$$

By the definition of of support, $\lambda(U) > 0$. Therefore,

$$
\Pr_\lambda(c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) \leq \varepsilon) = 1 - \Pr_\lambda(c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) > \varepsilon) \leq 1 - \lambda(U) < 1,
$$

which contradicts to that $\operatorname{ess\,sup}_\lambda c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) \leq \varepsilon$.

We next prove the "if" direction. By the law of total probability and the fact that $c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) \leq \varepsilon$ for any $(\mu,\Sigma) \in \mathrm{supp}(\lambda)$,

$$
\begin{aligned}
&\Pr_\lambda\left(c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) \leq \varepsilon\right) \\
=&\Pr_\lambda\left(c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) \leq \varepsilon | (\mu,\Sigma) \in \mathrm{supp}(\lambda)\right)\lambda(\mathrm{supp}(\lambda)) \\
&+ \Pr_\lambda\left(c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) \leq \varepsilon | (\mu,\Sigma) \notin \mathrm{supp}(\lambda)\right)(1-\lambda(\mathrm{supp}(\lambda))) \\
=&1 \cdot 1 + \Pr_\lambda\left(c((\mu,\Sigma),(\widehat{x},\sigma^2 I)) \leq \varepsilon | (\mu,\Sigma) \notin \mathrm{supp}(\lambda)\right) \cdot 0 = 1,
\end{aligned}
$$

which completes the proof. □

Proposition 5.4 (re-stated). *Fix any index $i \in \mathcal{I}_0$. For any $\widehat{x}_i \in \mathbb{R}^p$, $x \in \mathbb{R}^p$ and $\varepsilon_0 \in \mathbb{R}_+$, we have*

$$
\frac{\exp(-\alpha_i)}{(2\pi)^{p/2}} = \left\{\begin{array}{ll} \max & f(x|\mu_i,\Sigma_i) \\ \text{s.t.} & (\mu_i,\Sigma_i) \in \mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma} \\ & c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2 I)) \leq \varepsilon_0, \end{array}\right.
$$

where α_i is the optimal value of the two-dimensional optimization problem

$$
\min_{\substack{a\in\mathbb{R}_+,\ d_p\in[\sigma,+\infty) \\ a^2+(d_p-\sigma)^2\leq\varepsilon_0^2}} \log d_p + \frac{(\|x-\widehat{x}_i\|_2 - a)^2}{2d_p^2} + (p-1)\log\sigma.
$$

Proof of Proposition 5.4. Let α_i be the optimal value of the negative log-likelihood minimization problem

$$\alpha_i = \left\{ \begin{array}{ll} \min & \frac{1}{2}\log\det\Sigma_i + \frac{1}{2}(x-\mu_i)^\top\Sigma_i^{-1}(x-\mu_i) \\ \text{s.t.} & \mu_i \in \mathbb{R}^p,\ \Sigma_i \in \mathbb{S}_+^p \\ & \|\mu_i - \widehat{x}_i\|_2^2 + \mathrm{Tr}\left[\Sigma_i + \sigma^2 I - 2\big((\sigma^2 I)^{\frac{1}{2}}\Sigma_i(\sigma^2 I)^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \leq \varepsilon_0^2 \\ & \Sigma_i \succeq \sigma^2 I. \end{array} \right.$$

It is easy to see that

$$\max\{f(x|\mu_i,\Sigma_i) : (\mu_i,\Sigma_i) \in \mathbb{R}^p \times \mathbb{S}_{\geq\sigma}^p,\ c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2 I)) \leq \varepsilon_0\} = \frac{1}{\sqrt{(2\pi)^p}}\exp(-\alpha_i).$$

It remains to provide a simpler formulation to determine α_i. To simplify the notation, we omit the index i on all variables and parameters. We reparameterize $\Sigma = V\,\mathrm{diag}(d^2)V^\top$ for a vector $d \in \mathbb{R}_+^p$, where $\mathrm{diag}(d^2)$ denotes a $\mathbb{R}^{p\times p}$ diagonal matrix with its j-th diagonal entries equals to d_j^2, and $\mathrm{O}(p)$ is the set of p-dimensional orthogonal matrices

$$\mathrm{O}(p) = \{V \in \mathbb{R}^{p\times p} : V^\top V = I_p\}.$$

The negative log-likelihood minimization problem is further equivalent to

$$\begin{array}{ll} \min & \sum_{j=1}^p \log d_j + \frac{1}{2}(V^\top(x-\mu))^\top\,\mathrm{diag}(d^{-2})(V^\top(x-\mu)) \\ \text{s.t.} & d \in \mathbb{R}_+^p,\ V \in \mathrm{O}(p),\ \mu \in \mathbb{R}^p \\ & \|\mu - \widehat{x}\|_2^2 + \sum_{j=1}^p (d_j-\sigma)^2 \leq \varepsilon_0^2 \\ & d \geq \sigma, \end{array}$$

where $d \geq \sigma$ implies the element-wise constraints $d_j \geq \sigma$ for any $j = 1,\ldots,p$. We introduce an auxiliary variable $a \in \mathbb{R}_+$ and rewrite the optimization problem in an equivalent way as

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}_+^p,\ d\geq\sigma \\ a^2+\sum_j(d_j-\sigma)^2\leq\varepsilon_0^2}} \ \min_{\substack{\mu\in\mathbb{R}^p \\ \|\mu-\widehat{x}\|_2^2=a^2}} \ \min_{V\in\mathrm{O}(p)} \ \sum_{j=1}^p \log d_j + \frac{1}{2}(V^\top(x-\mu))^\top\,\mathrm{diag}(d^{-2})(V^\top(x-\mu)).$$

Notice that the above optimization problem is invariant to the ordering of the entries of d. As a consequence, without any loss of generality, we can assume that d_p is the maximum value across all d_j. By Lemma B.1, the above optimization problem becomes

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}_+^p,\ d\geq\sigma \\ a^2+\sum_j(d_j-\sigma)^2\leq\varepsilon_0^2 \\ d_p=\max\{d\}}} \ \min_{\substack{\mu\in\mathbb{R}^p \\ \|\mu-\widehat{x}\|_2^2=a^2}} \ \sum_{j=1}^p \log d_j + \frac{1}{2d_p^2}\|x-\mu\|_2^2.$$

Using Lemma B.2, we obtain the equivalent optimization problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}_+^p,\ d\geq\sigma \\ a^2+\sum_j(d_j-\sigma)^2\leq\varepsilon_0^2 \\ d_p=\max\{d\}}} \ \sum_{j=1}^p \log d_j + \frac{1}{2d_p^2}(\|x-\widehat{x}\|_2 - a)^2.$$

Rewriting the above problem into a two-layer optimization problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d_p\in\mathbb{R}_+,\ d_p\geq\sigma \\ a^2+(d_p-\sigma)^2\leq\varepsilon_0^2}} \left\{ \log d_p + \frac{1}{2d_p^2}(\|x-\widehat{x}\|_2 - a)^2 + \min_{\substack{d_j\in\mathbb{R}_+,\ d_j\geq\sigma\ \forall j=1,\ldots,p-1 \\ \sum_{j=1}^{p-1}(d_j-\sigma)^2\leq\varepsilon_0^2-a^2-(d_p-\sigma)^2 \\ d_j\leq d_p\ \forall j=1,\ldots,p-1}} \sum_{j=1}^{p-1}\log d_j \right\}. \tag{1}$$

Notice that for any d_p that is feasible for the outer minimization problem, the inner minimization problem over d_j, $\forall j = 1,\ldots,p-1$ admits a non-empty feasible set. Indeed, because $d_p \geq \sigma$, the value $d_j = \sigma$, $j = 1,\ldots,p-1$ is a

feasible solution for the inner problem. We now focus on solving the inner minimization problem. As $\log(\,\cdot\,)$ is an increasing function, for any $s \geq 0$, we find

$$\min_{\substack{d_p \geq d_j \geq \sigma \ \forall j=1,\ldots,p-1 \\ \sum_{j=1}^{p-1}(d_j-\sigma)^2 \leq s}} \sum_{j=1}^{p-1} \log d_j = (p-1)\log\sigma,$$

which holds because the optimization problem on the left hand side admits the optimal solution $d_j^\star = \sigma$ for all $j = 1, \ldots, p-1$. This completes the proof. □

Proposition 5.5 (re-stated). *Fix any index $i \in \mathcal{I}_1$. For any $\widehat{x}_i \in \mathbb{R}^p$, $x \in \mathbb{R}^p$ and $\varepsilon_1 \in \mathbb{R}_+$, we have*

$$\frac{\exp(\alpha_i)}{(2\pi)^{p/2}} = \left\{ \begin{array}{ll} \min & f(x|\mu_i, \Sigma_i) \\ \text{s.t.} & (\mu_i, \Sigma_i) \in \mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma} \\ & c((\mu_i, \Sigma_i), (\widehat{x}_i, \sigma^2 I)) \leq \varepsilon_1, \end{array} \right.$$

where α_i is the optimal value of the two-dimensional optimization problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d_1\in[\sigma,+\infty) \\ a^2+p(d_1-\sigma)^2\leq\varepsilon_1^2}} \left\{ -\log d_1 - \frac{(\|x-\widehat{x}_i\|_2 + a)^2}{2d_1^2} - (p-1)\log\left(\sigma + \sqrt{\frac{\varepsilon^2 - a^2 - (d_1-\sigma)^2}{p-1}}\right)\right\}.$$

Proof of Proposition 5.5. Let α_i be the optimal value of the log-likelihood *minimization* problem

$$\alpha_i = \left\{ \begin{array}{ll} \min & -\frac{1}{2}\log\det\Sigma_i - \frac{1}{2}(x-\mu_i)^\top\Sigma^{-1}(x-\mu_i) \\ \text{s.t.} & \mu_i \in \mathbb{R}^p,\ \Sigma_i \in \mathbb{S}^p_+ \\ & \|\mu_i - \widehat{x}_i\|_2^2 + \mathrm{Tr}\left[\Sigma_i + \sigma^2 I - 2\big((\sigma^2 I)^{\frac{1}{2}}\Sigma_i(\sigma^2 I)^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \leq \varepsilon_1^2 \\ & \Sigma_i \succeq \sigma^2 I. \end{array} \right.$$

It is easy to see that

$$\min\{f(x|\mu_i,\Sigma_i) : (\mu_i,\Sigma_i) \in \mathbb{R}^p \times \mathbb{S}^p_{\geq\sigma},\ c((\mu_i,\Sigma_i),(\widehat{x}_i,\sigma^2 I)) \leq \varepsilon_1\} = \frac{1}{(2\pi)^{p/2}}\exp(\alpha_i).$$

It remains to provide the computational routine to determine α_i. To simplify the notation, we omit the index i on all variables and parameters. We reparameterize $\Sigma = V\,\mathrm{diag}(d^2)V^\top$ for a vector $d \in \mathbb{R}^p_+$, where $\mathrm{diag}(d^2)$ denotes a $\mathbb{R}^{p\times p}$ diagonal matrix with its j-th diagonal entries equals to d_j^2, and $\mathrm{O}(p)$ is the set of p-dimensional orthogonal matrices

$$\mathrm{O}(p) = \{V \in \mathbb{R}^{p\times p} : V^\top V = I_p\}.$$

The log-likelihood minimization problem is further equivalent to

$$\begin{array}{ll} \min & -\sum_{j=1}^p \log d_j - \frac{1}{2}(V^\top(x-\mu))^\top\,\mathrm{diag}(d^{-2})(V^\top(x-\mu)) \\ \text{s.t.} & d \in \mathbb{R}^p_+,\ V \in \mathrm{O}(p),\ \mu \in \mathbb{R}^p \\ & \|\mu - \widehat{x}\|_2^2 + \sum_{j=1}^p (d_j-\sigma)^2 \leq \varepsilon_1^2 \\ & d \geq \sigma, \end{array}$$

where $d \geq \sigma$ implies the element-wise constraints $d_j \geq \sigma$ for any $j = 1, \ldots, p$. We introduce an auxiliary variable $a \in \mathbb{R}_+$ and rewrite the optimization problem in an equivalent way as

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}^p_+,\ d\geq\sigma \\ a^2+\sum_j(d_j-\sigma)^2\leq\varepsilon_1^2}} \ \min_{\substack{\mu\in\mathbb{R}^p \\ \|\mu-\widehat{x}\|_2^2=a^2}} \ \min_{V\in\mathrm{O}(p)} \ -\sum_{j=1}^p \log d_j - \frac{1}{2}(V^\top(x-\mu))^\top\,\mathrm{diag}(d^{-2})(V^\top(x-\mu)).$$

Notice that the above optimization problem is invariant to the ordering of the entries of d. As a consequence, without any loss of generality, we can assume that d_1 is the minimum value across all d_j. By Lemma B.1, the above optimization problem becomes

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}^p_+,\ d\geq\sigma \\ a^2+\sum_j(d_j-\sigma)^2\leq\varepsilon_1^2 \\ d_1=\min\{d\}}} \ \min_{\substack{\mu\in\mathbb{R}^p \\ \|\mu-\widehat{x}\|_2^2=a^2}} \ -\sum_{j=1}^p \log d_j - \frac{1}{2d_1^2}\|x-\mu\|_2^2.$$

Using Lemma B.2, we obtain the equivalent optimization problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d\in\mathbb{R}^p_+,\ d\geq\sigma\\ a^2+\sum_j(d_j-\sigma)^2\leq\varepsilon_1^2\\ d_1=\min\{d\}}} -\sum_{j=1}^p \log d_j - \frac{1}{2d_1^2}(\|x-\widehat{x}\|_2+a)^2.$$

Notice that the constraint $\sigma \leq d_1 = \min\{d\}$ implies that $p(d_1-\sigma)^2 \leq \sum_j (d_j-\sigma)^2$. As a consequence, any feasible value for d_1 should satisfy $a^2+p(d_1-\sigma)^2 \leq \varepsilon_1^2$. Separating the variable d into two groups d_1 and $d_2,\ldots,d_p$ leads to a two-layer optimization problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d_1\in\mathbb{R}_+,\ d_1\geq\sigma\\ a^2+p(d_1-\sigma)^2\leq\varepsilon_1^2}} \left\{ -\log d_1 - \frac{1}{2d_1^2}(\|x-\widehat{x}\|_2+a)^2 + \min_{\substack{d_j\in\mathbb{R}_+,\ d_j\geq d_1\ \forall j=2,\ldots,p\\ \sum_{j=2}^p (d_j-\sigma)^2\leq\varepsilon_1^2-a^2-(d_1-\sigma)^2}} -\sum_{j=2}^p \log d_j \right\}. \tag{2}$$

Consider momentarily the minimization problem

$$\min_{\substack{d_j\in\mathbb{R}_+\ \ \forall j=2,\ldots,p\\ \sum_{j=2}^p (d_j-\sigma)^2\leq\varepsilon_1^2-a^2-(d_1-\sigma)^2}} -\sum_{j=2}^p \log d_j,$$

where the constraints $d_j \geq d_1$ have been intentionally omitted. Proposition B.3 asserts that this optimization problem has the optimal value

$$-(p-1)\log\left(\sigma+\sqrt{\frac{\varepsilon_1^2-a^2-(d_1-\sigma)^2}{p-1}}\right)$$

at the optimal solution $d_j^\star = \sigma + \sqrt{\frac{\varepsilon_1^2-a^2-(d_1-\sigma)^2}{p-1}}$, which also by the outer constraint $a^2+p(d_1-\sigma)^2 \leq \varepsilon_1^2$ satisfies $d_j \geq d_1\ \forall j=2,\ldots,p$. Thus it is indeed the optimal solution to the inner minimization problem in (2). As a consequence, problem (2) is equivalent to

$$\min_{\substack{a\in\mathbb{R}_+,\ d_1\in\mathbb{R}_+,\ d_1\geq\sigma\\ a^2+p(d_1-\sigma)^2\leq\varepsilon_1^2}} \left\{ -\log d_1 - \frac{1}{2d_1^2}(\|x-\widehat{x}\|_2+a)^2 - (p-1)\log\left(\sigma+\sqrt{\frac{\varepsilon_1^2-a^2-(d_1-\sigma)^2}{p-1}}\right) \right\}.$$

This completes the proof. □

B AUXILIARY RESULTS

The following preparatory results are necessary to prove Propositions 5.4 and 5.5.

Lemma B.1 (Eigenbasis solution)**.** *Let $E \in \mathbb{R}^{p\times p}$ be a diagonal matrix satisfying $E_{11} \leq \cdots \leq E_{pp}$. Then, for any $w \in \mathbb{R}^p$, we have*

$$\max_{V\in \mathrm{O}(p)} w^\top V E V^\top w = E_{pp}\|w\|_2^2.$$

Proof of Lemma B.1. The claim holds trivially when $w=0$. Consider now any $w \in \mathbb{R}^p\backslash\{0\}$. Since $VEV^\top \preceq E_{pp}\cdot I_p$, we find

$$\max_{V\in \mathrm{O}(p)} w^\top V E V^\top w \leq \max_{V\in \mathrm{O}(p)} w^\top V (E_{pp}\cdot I_p) V^\top w = E_{pp}\|w\|_2^2.$$

On the other hand, taking $V^\star = [v_1^\star,\ldots,v_p^\star] \in \mathrm{O}(p)$ with $v_p^\star = \frac{w}{\|w\|_2}$, and using the orthogonality of the columns of $V^\star$, we have

$$w^\top V^\star E {V^\star}^\top w = E_{pp}\|w\|_2^2.$$

This shows that $V^\star$ is an optimal solution and completes the proof. □

Lemma B.2 (Quadratic optimization). *For any $x \in \mathbb{R}^p$, $\widehat{x} \in \mathbb{R}^p$ and $a \in \mathbb{R}_+$, the following assertions hold.*

- *Convex quadratic minimization:*

$$\min_{\mu \in \mathbb{R}^p : \|\mu - \widehat{x}\|_2^2 = a^2} \|x - \mu\|_2^2 = (\|x - \widehat{x}\|_2 - a)^2,$$

  where the minimum is attained at $\mu^\star = \frac{a}{\|x-\widehat{x}\|_2} x + (1 - \frac{a}{\|x-\widehat{x}\|_2})\widehat{x}$.
- *Convex quadratic maximization:*

$$\max_{\mu \in \mathbb{R}^p : \|\mu - \widehat{x}\|_2^2 = a^2} \|x - \mu\|_2^2 = (\|x - \widehat{x}\|_2 + a)^2,$$

  where the maximum is attained at $\mu^\star = -\frac{a}{\|x-\widehat{x}\|_2} x + (1 + \frac{a}{\|x-\widehat{x}\|_2})\widehat{x}$.

The results in Lemma B.2 are dispersed in the literature. An elementary proof is provided here for completeness.

Proof of Lemma B.2. By the triangle inequality, for any μ such that $\|\mu - \widehat{x}\|_2 = a$, we have

$$\|x - \mu\|_2 \geq |\|x - \widehat{x}\| - \|\mu - \widehat{x}\|| = |\|x - \widehat{x}\| - a|,$$

where the lower bound can be attained by taking $\mu = \frac{a}{\|x-\widehat{x}\|_2} x + (1 - \frac{a}{\|x-\widehat{x}\|_2})\widehat{x}$. Therefore,

$$\min_{\mu \in \mathbb{R}^p : \|\mu - \widehat{x}\|_2^2 = a^2} \|x - \mu\|_2^2 = (\|x - \widehat{x}\|_2 - a)^2$$

Similarly, by the triangle inequality we have

$$\|x - \mu\|_2 \leq \|x - \widehat{x}\| + \|\widehat{x} - \mu\| = \|x - \widehat{x}\| + a,$$

and the upper bound can be attained by $\mu = -\frac{a}{\|x-\widehat{x}\|_2} x + (1 + \frac{a}{\|x-\widehat{x}\|_2})\widehat{x}$. This completes the proof. □

Proposition B.3 (Logarithm maximization). *For any $s, \sigma \geq 0$ and positive integer k, we have*

$$k \log\left(\sqrt{\frac{s}{k}} + \sigma\right) = \left\{ \begin{array}{ll} \max\limits_{e \in \mathbb{R}_+^k} & \sum\limits_{j=1}^k \log e_j \\ \text{s. t.} & \sum\limits_{j=1}^k (\sigma - e_j)^2 \leq s. \end{array} \right. \tag{3}$$

Moreover, the optimal solution $e^\star$ satisfies $e_j^\star = \sqrt{\frac{s}{k}} + \sigma$ for any $j = 1, \ldots, k$.

Proof of Proposition B.3. Let $e^\star \in \mathbb{R}_+^k$ be an optimal solution to the maximization problem (3). Suppose there exist two indices m and n such that $e_m^\star \neq e_n^\star$. Consider e' defined by

$$e_j' = \left\{ \begin{array}{ll} \frac{1}{2}(e_m^\star + e_n^\star), & \text{if } j \in \{m, n\}, \\ e_j^\star, & \text{otherwise.} \end{array} \right.$$

By the convexity of the function $x \mapsto (x - \sigma)^2$,

$$(e_m' - \sigma)^2 + (e_n' - \sigma)^2 = 2\left(\frac{e_m^\star + e_n^\star}{2} - \sigma\right)^2 \leq (e_m^\star - \sigma)^2 + (e_n^\star - \sigma)^2,$$

which implies that e' is a feasible solution to problem (3). Furthermore, since $e_m^\star \neq e_n^\star$, by the concavity of the function $x \mapsto \log x$, we have that

$$\log e_m^\star + \log e_n^\star < 2 \log\left(\frac{e_m^\star + e_n^\star}{2}\right) = \log e_m' + \log e_n',$$

which violates the optimality of $e^\star$. Therefore, any optimal solution $e^\star$ must have all entries identical. Using this, we get from the constraint that

$$|e_j^\star - \sigma| \leq \sqrt{\frac{s}{k}} \quad \forall j = 1, \ldots, k.$$

By continuity of the objective and constraint functions, we must have

$$|e_j^\star - \sigma| = \sqrt{\frac{s}{k}} \quad \forall j = 1, \ldots, k.$$

Since the objective function is increasing in $e_j^\star$, the optimal solution is given by

$$e_j^\star = \sigma + \sqrt{\frac{s}{k}} \quad \forall j = 1, \ldots, k.$$

The optimal value can then be obtained by direct computation. This completes the proof. □

C FIRST-ORDER ALGORITHMS

C.1 OPTIMISTIC LIKELIHOOD PROBLEM

For the optimistic likelihood problem, Theorem 5.1 reduces the task to solving the 2-dimensional problem

$$\min_{\substack{a \in \mathbb{R}_+,\ d_p \in [\sigma, +\infty) \\ a^2 + (d_p - \sigma)^2 \leq \varepsilon_0^2}} \log d_p + \frac{(\|x - \widehat{x}_i\|_2 - a)^2}{2d_p^2} + (p-1)\log\sigma.$$

By letting

$$d_p = v_2 + \sigma, \quad \text{and} \quad a = v_1,$$

we can obtain the equivalent form

$$\min_{\substack{v_1, v_2 \geq 0 \\ v_1^2 + v_2^2 \leq \varepsilon_0^2}} F(v), \tag{4}$$

where the objective function is given by

$$F(v) = \log(v_2 + \sigma) + \frac{(\|x - \widehat{x}_i\|_2 - v_1)^2}{2(v_2 + \sigma)^2} + (p-1)\log\sigma.$$

If we denote by $\mathcal{V} = \{v \in \mathbb{R}^2 : v_1, v_2 \geq 0, v_1^2 + v_2^2 \leq \varepsilon_0^2\}$ the feasible region of the above minimization problem, then the projection $\mathrm{Proj}_{\mathcal{V}}(v)$ can be computed in closed-form via

$$\mathrm{Proj}_{\mathcal{V}}(v) = \begin{cases} v, & \text{if } v_1, v_2 \geq 0, v_1^2 + v_2^2 \leq \varepsilon_0^2, \\ \frac{\varepsilon_0}{\|v\|_2} v, & \text{if } v_1, v_2 \geq 0, v_1^2 + v_2^2 > \varepsilon_0^2, \\ (0, \varepsilon_0)^\top, & \text{if } v_1 < 0, v_2 > \varepsilon_0, \\ (0, v_2)^\top, & \text{if } v_1 < 0, 0 \leq v_2 \leq \varepsilon_0, \\ (\varepsilon_0, 0)^\top, & \text{if } v_1 > \varepsilon_0, v_2 < 0, \\ (v_1, 0)^\top, & \text{if } 0 \leq v_1 \leq \varepsilon_0, v_2 < 0, \\ (0, 0)^\top, & \text{if } v_1, v_2 < 0. \end{cases}$$

Algorithm 1 is a projected gradient descent routine to solve problem (4). The convergence guarantee for Algorithm 1 follows from Beck [2017, Theorem 10.15].

Algorithm 1 Projected Gradient Descent Algorithm with Backtracking Line-Search

Algorithm parameters: Line search parameters $\theta \in (0,1)$, $\beta > 0$
Initialization: Set $v^0 \leftarrow 0$
for $t = 0, 1, \ldots$ **do**
Find the smallest integer $k \geq 0$ such that

$$F\left(\mathrm{Proj}_{\mathcal{V}}(v^t - \theta^k \beta \nabla F(v^t))\right) \leq F(v^t) - \frac{1}{2\theta^k \beta}\|v^t - \mathrm{Proj}_{\mathcal{V}}(v^t - \theta^k \beta \nabla F(v^t))\|_2^2$$

Set $s^t = \theta^k \beta$ and set $v^{t+1} = \mathrm{Proj}_{\mathcal{V}}(u^t - s^t \nabla F(v^t))$.
end for

C.2 PESSIMISTIC LIKELIHOOD PROBLEM

For the pessimistic likelihood problem, Theorem 5.2 reduces the task to solving the 2-dimensional problem

$$\min_{\substack{a\in\mathbb{R}_+,\ d_1\in[\sigma,+\infty)\\ a^2+p(d_1-\sigma)^2\leq\varepsilon_1^2}} \left\{ -\log d_1 - \frac{1}{2d_1^2}(\|x-\widehat{x}_i\|_2 + a)^2 - (p-1)\log\left(\sigma + \sqrt{\frac{\varepsilon_1^2 - a^2 - (d_1-\sigma)^2}{p-1}}\right)\right\}.$$

Note that the gradient of the objective function is a non-Lipschitz function. Worse still, the gradient is even undefined on at the feasible point $(d_1, a) = (\sigma, \varepsilon_1)$. These properties induce numerical issues for the optimization algorithm. Therefore, we solve the following perturbed problem instead:

$$\min_{\substack{a\in\mathbb{R}_+,\ d_1\in[\sigma,+\infty)\\ a^2+p(d_1-\sigma)^2\leq\varepsilon_1^2}} \left\{ -\log d_1 - \frac{1}{2d_1^2}(\|x-\widehat{\mu}\|_2 + a)^2 - (p-1)\log\left(\sigma + \sqrt{\frac{\zeta+\varepsilon_1^2 - a^2 - (d_1-\sigma)^2}{p-1}}\right)\right\}, \tag{5}$$

for some small $\zeta > 0$. By Bonnans and Shapiro [2013, Proposition 4.4], the optimal value of problem (5) is continuous in ζ and the optimal solution set is upper semi-continuous in ζ as a set-valued mapping, see Bonnans and Shapiro [2013, Section 4.1].

We now derive a projected gradient descent algorithm with backtracking line search for solving problem (5). First, by letting

$$d_1 = u_2 + \sigma, \quad \text{and} \quad a = \sqrt{p}u_1,$$

we can equivalently transform problem (5) to the following one:

$$\min_{\substack{u_1,u_2\geq 0\\ u_1^2+u_2^2\leq(\varepsilon_1/\sqrt{p})^2}} F(u), \tag{6}$$

where the objective function is given by

$$F(u) = -\log(u_2+\sigma) - \frac{1}{2(u_2+\sigma)^2}(\|x-\widehat{x}_i\|_2 + \sqrt{p}u_1)^2 - (p-1)\log\left(\sigma + \sqrt{\frac{\zeta+\varepsilon_1^2 - pu_1^2 - u_2^2}{p-1}}\right).$$

The upshot of problem (6) is that the feasible region is the intersection of the non-negative orthant with a circular disk of radius $\varepsilon_1/\sqrt{p}$ centered at the origin. As we will see below, this enables easy computation of the projection and linear optimization oracle. Indeed, denoting by $\mathcal{U} = \{u \in \mathbb{R}^2 : u_1, u_2 \geq 0, u_1^2 + u_2^2 \leq (\varepsilon_1/\sqrt{p})^2\}$ the feasible region of problem (6), the projection $\mathrm{Proj}_{\mathcal{U}}(u)$ can be computed in closed-form via

$$\mathrm{Proj}_{\mathcal{U}}(u) = \begin{cases} u, & \text{if } u_1, u_2 \geq 0, u_1^2 + u_2^2 \leq (\varepsilon_1/\sqrt{p})^2, \\ \frac{(\varepsilon_1/\sqrt{p})}{\|u\|_2}u, & \text{if } u_1, u_2 \geq 0, u_1^2 + u_2^2 > (\varepsilon_1/\sqrt{p})^2, \\ (0, \frac{\varepsilon_1}{\sqrt{p}})^\top, & \text{if } u_1 < 0, u_2 > \frac{\varepsilon_1}{\sqrt{p}}, \\ (0, u_2)^\top, & \text{if } u_1 < 0, 0 \leq u_2 \leq \frac{\varepsilon_1}{\sqrt{p}}, \\ (\frac{\varepsilon_1}{\sqrt{p}}, 0)^\top, & \text{if } u_1 > \frac{\varepsilon_1}{\sqrt{p}}, u_2 < 0, \\ (u_1, 0)^\top, & \text{if } 0 \leq u_1 \leq \frac{\varepsilon_1}{\sqrt{p}}, u_2 < 0, \\ (0,0)^\top, & \text{if } u_1, u_2 < 0. \end{cases}$$

A projected gradient descent algorithm can now be employed to solve problem (6).

D RECOVERY OF THE ADVERSARIAL DISTRIBUTION

It is often instructive to recover and analyze the optimal distribution that maximizes the posterior probability odds ratio, or more directly, the likelihood ratio in (2). Equivalent, it suffices to characterize the distribution $\mathbb{Q}_0^\star$ that maximizes (3), and the distribution $\mathbb{Q}_1^\star$ that minimizes (4).

Lemma D.1 (Likelihood maximizer)**.** *For each $i \in \mathcal{I}_0$, let $(a_i^\star, d_{pi}^\star)$ be the optimal solution of the following two-dimensional optimization problem*

$$\min_{\substack{a\in\mathbb{R}_+,\ d_p\in[\sigma,+\infty)\\ a^2+(d_p-\sigma)^2\leq\varepsilon_0^2}} \log d_p + \frac{(\|x-\widehat{x}_i\|_2 - a)^2}{2d_p^2} + (p-1)\log\sigma.$$

Then, the maximizero $\mathbb{Q}_0^\star$ of problem (3) *is a Gaussian mixture with N_0 components, and for $i \in \mathcal{I}_0$, the i-th components has mean*

$$\mu_i^\star = \frac{a_i^\star}{\|x-\widehat{x}_i\|_2} x + \left(1 - \frac{a_i^\star}{\|x-\widehat{x}_i\|_2}\right)\widehat{x}_i,$$

and covariance matrix

$$\Sigma_i^\star = V_i^\star \operatorname{diag}(\sigma,\ldots,\sigma,d_{pi}^\star)^2 (V_i^\star)^\top,$$

where $V_i^\star$ is any orthogonal matrix with the p-th column given by $\frac{x-\mu_i^\star}{\|x-\mu_i^\star\|_2}$.

Proof of Lemma D.1. The result follows directly by inspecting the proofs of Proposition 5.4, Lemma B.1 and Lemma B.2. □

Lemma D.2 (Likelihood minimizer)**.** *For each $i \in \mathcal{I}_1$, let $(a_i^\star, d_{1i}^\star)$ be the optimal solution of the following two-dimensional optimization problem*

$$\min_{\substack{a\in\mathbb{R}_+,\ d_1\in[\sigma,+\infty)\\ a^2+p(d_1-\sigma)^2\leq\varepsilon_1^2}} \left\{ -\log d_1 - \frac{(\|x-\widehat{x}_i\|_2 + a)^2}{2d_1^2} - (p-1)\log\left(\sigma + \sqrt{\frac{\varepsilon_1^2 - a^2 - (d_1-\sigma)^2}{p-1}}\right)\right\}.$$

Then, the minimizer $\mathbb{Q}_1^\star$ of problem (4) *is a Gaussian mixture with N_1 components, and for $i \in \mathcal{I}_1$, the i-th components has mean*

$$\mu_i^\star = -\frac{a_i^\star}{\|x-\widehat{x}_i\|_2} x + \left(1 + \frac{a_i^\star}{\|x-\widehat{x}_i\|_2}\right)\widehat{x}_i,$$

and covariance matrix

$$\Sigma_i^\star = V_i^\star \operatorname{diag}\left(d_{1i}^\star, \sigma + \sqrt{\frac{\varepsilon_1^2 - {a_i^\star}^2 - (d_{1i}^\star-\sigma)^2}{p-1}},\ldots,\sigma + \sqrt{\frac{\varepsilon_1^2 - {a_i^\star}^2 - (d_{1i}^\star-\sigma)^2}{p-1}}\right)^2 (V_i^\star)^\top,$$

where $V_i^\star$ is any orthogonal matrix with the 1st column given by $\frac{x-\mu_i^\star}{\|x-\mu_i^\star\|_2}$.

Proof of Lemma D.2. The result follows directly by inspecting the proofs of Proposition 5.5, Lemma B.1 and Lemma B.2. □

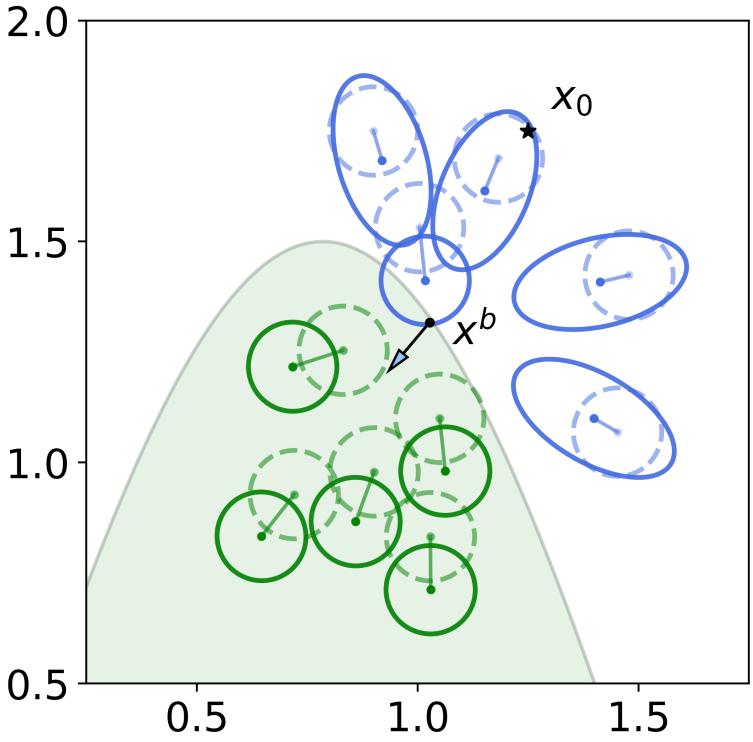

Figure 1: Visualization of the worst-case distributions on a toy dataset, color codes are similar to Figure 1. The dashed, opaque dots and circles represent the isotropic Gaussian around each data sample. The solid dots and circles represent the worst-case distributions corresponding to the boundary point x^b. For blue (unfavorably predicted) samples, the worst-case distribution is formed by perturbing the distribution towards x^b – which leads to maximizing the posterior probability of unfavorable prediction. For green (favorably predicted) samples, the worst-case distribution is formed by perturbing the distribution away from x^b – which leads to minimizing the posterior probability of favorable prediction. These worst-case distributions will maximize the posterior probability odds ratio.

References

Amir Beck. *First-order Methods in Optimization*. SIAM, 2017.

J Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Science & Business Media, 2013.