A Codes for numerical experiment

All codes for the numerical experiment can be found in `https://github.com/ai-submit/OptimalWasserstein`.

B Additional numerical experiment

B.1 PDE-constrained linear Bayesian inference

In this experiment, we consider a linear Bayesian inference problem constrained by a partial differential equation (PDE) model for contaminant diffusion in environmental engineering in domain $D = (0, 1)$,

$$-\kappa \Delta u + \nu u = x \quad \text{in } D,$$

where x is a contaminant source field parameter in domain D, u is the contaminant concentration which we can observe at some locations, κ and ν are diffusion and reaction coefficients. For simplicity, we set $\kappa, \nu = 1$, $u(0) = u(1) = 0$, and consider 15 pointwise observations of u with 1% noise, equidistantly distributed in D. We consider a Gaussian prior distribution $x \sim \mathcal{N}(0, C)$ with covariance given by a differential operator $C = (-\delta\Delta + \gamma I)^{-\alpha}$ with $\delta, \gamma, \alpha > 0$ representing the correlation length and variance, which is commonly used in geoscience. We set $\delta = 0.1, \gamma = 1, \alpha = 1$. In this linear setting, the posterior is Gaussian with the mean and covariance given analytically, which are used as reference to assess the sample goodness. We solve this forward model by a finite element method with piece-wise linear elements on a uniform mesh of size 2^k, $k \geq 1$. We project this high-dimensional parameter to the data-informed low dimensions as in Wang et al. (2021) to alleviate the curse of dimensionality when applying WGD-cvxNN and WGD-NN, which we call pWGD-cvxNN and pWGD-NN, respectively. For $k = 4$ we have 17 dimensions for the discrete parameter and 4 dimensions after projection.

We run pWGD-cvxNN and pWGD-NN using 16 samples for 200 iterations with $\alpha_l = 10^{-3}$, $\beta = 5$, $\gamma_1 = 0.95$, and $\gamma_2 = 0.95^{10}$ for both methods. We use $m = 200$ neurons for pWGD-NN and train it by the Adam optimizer for 200 sub-iterations as in the first example. From Figure 5, we observe that pWGD-cvxNN achieves better root mean squared error (RMSE) than pWGD-NN for both the sample mean and the sample variance compared to the reference.

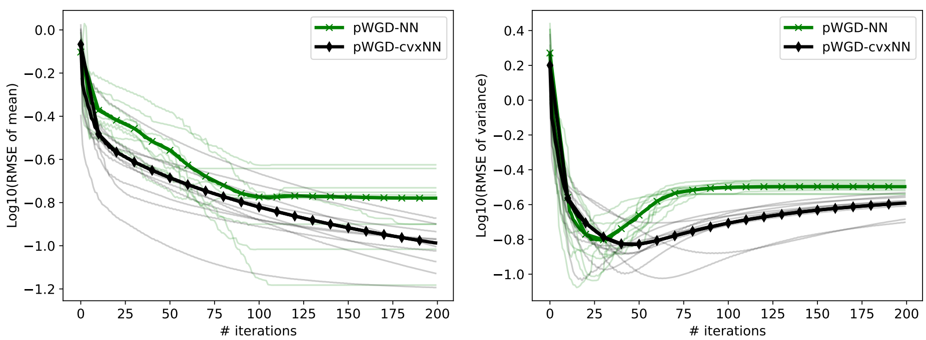

Figure 5: Ten trials and the RMSE of the sample mean (top) and sample variance (bottom) by pWGD-NN and pWGD-cvxNN at different iterations. Linear inference problem.

C Choice of the regularization parameter

As the constraints in the relaxed dual problem (16) depends on the regularization parameter $\tilde{\beta}$, it is possible that for small $\tilde{\beta}$, the relaxed dual problem (16) is infeasible. Consider the following SDP

$$
\begin{aligned}
\min\ \tilde{\beta},\ \text{s.t. } & \tilde{A}_j(\Lambda) + \tilde{B}_j + \sum_{n=0}^{N} r_n^{(j,-)} H_n^{(j)} + \tilde{\beta} e_{d+1} e_{d+1}^T \succeq 0, \\
& -\tilde{A}_j(\Lambda) - \tilde{B}_j + \sum_{n=0}^{N} r_n^{(j,+)} H_n^{(j)} + \tilde{\beta} e_{d+1} e_{d+1}^T \succeq 0, \\
& r^{(j,-)} \geq 0, r^{(j,+)} \geq 0, j \in [p].
\end{aligned}
\tag{21}
$$

Here the variables are $\tilde{\beta}$, Λ and $\{r^{(j,+)}, r^{(j,-)}\}_{j=1}^p$. Let $\tilde{\beta}_1$ be the optimal value of the above problem. Then, only for $\tilde{\beta} \geq \tilde{\beta}_1$, there exists $\Lambda \in \mathbb{R}^{N\times d}$ satisfying the constraints in (16). In other words, the relaxed dual problem (16) is feasible. We also note that $\tilde{\beta}_1$ only depends on the samples X and it does not depend on the value of $\nabla \log \pi$ evaluated on $x_1, \dots, x_N$. On the other hand, consider the following SDP

$$
\begin{aligned}
\min\ \tilde{\beta},\ \text{s.t. } & \tilde{A}_j(Y) + \tilde{B}_j + \sum_{n=0}^{N} r_n^{(j,-)} H_n^{(j)} + \tilde{\beta} e_{d+1} e_{d+1}^T \succeq 0, \\
& -\tilde{A}_j(Y) - \tilde{B}_j + \sum_{n=0}^{N} r_n^{(j,+)} H_n^{(j)} + \tilde{\beta} e_{d+1} e_{d+1}^T \succeq 0, \\
& r^{(j,-)} \geq 0, r^{(j,+)} \geq 0, j \in [p],
\end{aligned}
\tag{22}
$$

where the variables are $\tilde{\beta}$ and $\{r^{(j,+)}, r^{(j,-)}\}_{j=1}^p$. Let $\tilde{\beta}_2$ be the optimal value of the above problem. For $\tilde{\beta} \geq \tilde{\beta}_2$, as $\mathbf{Y}$ is feasible for the constraints in (16), the optimal value of the relaxed dual problem (16) is 0. In short, only when $\tilde{\beta} \in [\tilde{\beta}_1, \tilde{\beta}_2]$, the variational problem (16) is non-trivial. To ensure that solving the relaxed dual problem (16) gives a good approximation of the Wasserstein gradient direction, we shall avoid choosing $\tilde{\beta}$ either too small or too large.

D Proofs

D.1 Proof of Proposition 1

PROOF We first note that

$$
\begin{aligned}
& \frac{1}{2}\int \|\nabla\Phi - \nabla \log \rho + \nabla \log \pi\|_2^2 \rho dx \\
=& \frac{1}{2}\int \|\nabla\Phi\|_2^2 \rho dx + \int \langle \nabla \log \pi - \nabla \log \rho, \nabla \Phi\rangle \rho dx \\
& + \frac{1}{2}\int \|\nabla \log \rho - \nabla \log \pi\|_2^2 \rho dx.
\end{aligned}
\tag{23}
$$

We notice that the term $\frac{1}{2}\int \|\nabla \log \rho - \nabla \log \pi\|_2^2 \rho dx$ does not depend on Φ. Utilizing the integration by parts, we can compute that

$$
\begin{aligned}
\int \langle \nabla \log \rho, \nabla \Phi \rangle \rho dx =& \int \left\langle \frac{\nabla \rho}{\rho}, \nabla \Phi \right\rangle \rho dx \\
=& \int \langle \nabla \rho, \nabla \Phi \rangle dx \\
=& -\int \Delta \Phi \rho dx.
\end{aligned}
\tag{24}
$$

Therefore, the variational problem (4) is equivalent to

$$
\inf_{\Phi \in C^\infty(\mathbb{R}^d)} \frac{1}{2}\int \|\nabla \Phi\|_2^2 \rho dx + \int \langle \nabla \log \pi, \nabla \Phi \rangle \rho dx + \int \Delta \Phi \rho dx. \tag{25}
$$

By restricting the domain $C^\infty(\mathbb{R}^d)$ to $\mathcal{H}$, we complete the proof.

D.2 Proof of Proposition 2

PROOF Suppose that $\hat{w}_i = \beta_i^{-1} w_i$ and $\hat{\alpha}_i = \beta_i^2 \alpha_i$, where $\beta_i > 0$ is a scale parameter for $i \in [m]$. Let $\boldsymbol{\theta}' = \{(\hat{w}_i, \hat{\alpha}_i)\}_{i=1}^m$. We note that

$$\hat{\alpha}_i \hat{w}_i \psi'(\hat{w}_i^T x_n) = \beta_i \alpha_i w_i \psi'\left(\beta_i^{-1} w_i^T x_n\right) = \alpha_i w_i \psi'(w_i^T x_n), \tag{26}$$

and

$$\hat{\alpha}_i \|\hat{w}_i\|_2^2 \psi''(\hat{w}_i^T x_n) = \alpha_i \|w_i\|_2^2 \psi''(\hat{w}_i^T x_n) = \alpha_i \|w_i\|_2^2 \psi''(w_i^T x_n). \tag{27}$$

This implies that $\Phi_{\boldsymbol{\theta}}(x) = \Phi_{\boldsymbol{\theta}'}(x)$ and $\nabla \cdot \Phi_{\boldsymbol{\theta}}(x) = \nabla \cdot \Phi_{\boldsymbol{\theta}'}(x)$. For the regularization term $R(\boldsymbol{\theta})$, we note that

$$\begin{aligned}
\|\hat{w}_i\|_2^3 + \|\hat{\alpha}_i\|_2^3 =& \beta_i^6 |\alpha_i|^3 + \beta_i^{-3} \|w_i\|_2^3 \\
=& \beta_i^6 |\alpha_i|^3 + \frac{1}{2}\beta_i^{-3} \|w_i\|_2^3 + \frac{1}{2}\beta_i^{-3} \|w_i\|_2^3 \\
=& 3 \cdot 2^{-2/3} \|w_i\|_2^2 |\alpha_i|.
\end{aligned} \tag{28}$$

The optimal scaling parameter is given by $\alpha_i = 2^{-1/9} \frac{\|w_i\|_2^{1/3}}{|\alpha_i|_1^{1/3}}$. As the scaling operation does not change $\|w_i\|_2^2 |\alpha_i|$, we can simply let $\|w_i\|_2 = 1$. Thus, the regularization term $\frac{\beta}{2} R(\boldsymbol{\theta})$ becomes $\frac{\tilde{\beta}}{N} \sum_{i=1}^m \|u_i\|_1$. This completes the proof.

D.3 Proof of Proposition 3

PROOF Consider the Lagrangian function

$$\begin{aligned}
L(Z, W, \alpha, \Lambda) =& \frac{1}{2}\|Z\|_F^2 + \sum_{n=1}^N \sum_{i=1}^m \alpha_i \|w_i\|_2^2 \psi''(w_i^T x_n) + \mathrm{tr}(Y^T Z) + \tilde{\beta}\|\alpha\|_1 \\
&+ \sum_{n=1}^N \lambda_n^T \left(z_n - \sum_{i=1}^m \alpha_i w_i \psi'(x_n^T w_i) \right) \\
=& \tilde{\beta}\|\alpha\|_1 + \sum_{i=1}^m \alpha_i \sum_{n=1}^N \left(\|w_i\|_2^2 \psi''(w_i^T x_n) - \lambda_n^T w_i \psi'(x_m^T w_i) \right) \\
&+ \frac{1}{2}\|Z\|_F^2 + \mathrm{tr}((Y + \Lambda)^T Z).
\end{aligned} \tag{29}$$

For fixed W, the constraints on Z and α are linear and the strong duality holds. Thus, we can exchange the order of $\min_{Z,\alpha}$ and $\max_\Lambda$. Thus, we can compute that

$$\begin{aligned}
&\min_{Z,W,\alpha} \max_\Lambda L(Z, W, \alpha, \Lambda) \\
=& \min_W \max_\Lambda \min_{\alpha, Z} L(Z, W, \alpha, \Lambda) \\
=& \min_W \max_\Lambda \min_{\alpha, Z} \tilde{\beta}\|\alpha\|_1 + \sum_{i=1}^m \alpha_i \sum_{n=1}^N \left(\|w_i\|_2^2 \psi''(w_i^T x_n) - \lambda_n^T w_i \psi'(x_m^T w_i) \right) + \frac{1}{2}\|Z\|_F^2 + \mathrm{tr}((Y + \Lambda)^T Z) \\
=& \min_W \max_\Lambda -\frac{1}{2}\|\Lambda + Y\|_F^2 + \sum_{i=1}^m \mathbb{I}\left(\max_{w_i : \|w_i\|_2 \le 1} \left| \sum_{n=1}^N \|w_i\|_2^2 \psi''(w_i^T x_n) - y_n^T w_i \psi'(x_n^T w_i) \right| \le \tilde{\beta} \right).
\end{aligned} \tag{30}$$

By exchanging the order of min and max, we can derive the dual problem:

$$\begin{aligned}
&\max_\Lambda \min_W -\frac{1}{2}\|\Lambda + Y\|_F^2 + \sum_{i=1}^m \mathbb{I}\left(\max_{w_i : \|w_i\|_2 \le 1} \left| \sum_{n=1}^N \|w_i\|_2^2 \psi''(w_i^T x_n) - y_n^T w_i \psi'(x_n^T w_i) \right| \le \tilde{\beta} \right) \\
=& \max_\Lambda -\frac{1}{2}\|\Lambda + Y\|_F^2 \text{ s.t. } \max_{w_i : \|w_i\|_2 \le 1} \left| \sum_{n=1}^N \|w_i\|_2^2 \psi''(w_i^T x_n) - y_n^T w_i \psi'(x_n^T w_i) \right| \le \tilde{\beta}, i \in [m] \\
=& \max_\Lambda -\frac{1}{2}\|\Lambda + Y\|_F^2 \text{ s.t. } \max_{w : \|w\|_2 \le 1} \left| \sum_{n=1}^N \|w\|_2^2 \psi''(w^T x_n) - y_n^T w \psi'(x_n^T w) \right| \le \tilde{\beta}, i \in [m]
\end{aligned} \tag{31}$$

This completes the proof.

D.4 Proof of Proposition 4

PROOF Based on the hyper-plane arrangements $D_1, \dots, D_p$, the dual constraint is equivalent to that for all $j \in [p]$,

$$\left|2\operatorname{tr}(D_j)\|w\|_2^2 - 2w^T\Lambda^T D_j X w\right| \le \tilde{\beta} \tag{32}$$

holds for all $w \in \mathbb{R}^d$ satisfying $\|w\|_2 \le 1, (2D_j - I)Xw \ge 0$. This is equivalent to say that for all $j \in [p]$

$$\begin{aligned}
-\tilde{\beta} \ge& \min 2\operatorname{tr}(D_j)\|w\|_2^2 - 2w^T\Lambda^T D_j X w,\\
&\text{s.t. } \|w\|_2 \le 1, 2(D_j - I)Xw \ge 0,\\
\tilde{\beta} \le& \max 2\operatorname{tr}(D_j)\|w\|_2^2 - 2w^T\Lambda^T D_j X w,\\
&\text{s.t. } \|w\|_2 \le 1, 2(D_j - I)Xw \ge 0.
\end{aligned} \tag{33}$$

From a convex optimization perspective, the natural idea to interpret the constraint (33) is to transform the minimization problem into a maximization problem. We can rewrite the minimization problem in (33) as a trust region problem with inequality constraints:

$$\begin{aligned}
\min_{w \in \mathbb{R}^d} \; & w^T\left(B_j + A_j(\Lambda)\right)w,\\
\text{s.t. } & \|w\|_2 \le 1, (2D_j - I)Xw \ge 0.
\end{aligned} \tag{34}$$

As the problem (34) is a convex problem, by taking the dual of (34) w.r.t. w, we can transform (34) into a maximization problem. However, as (34) is a trust region problem with inequality constraints, the dual problem of (34) can be very complicated. According to (Jeyakumar & Li, 2014), the optimal value of the problem (34) is bounded by the optimal value of the following SDP

$$\begin{aligned}
\min_{Z \in \mathbb{S}^{d+1}} \; & \operatorname{tr}((\tilde{A}_j(\Lambda) + \tilde{B}_j)Z),\\
\text{s.t. } & \operatorname{tr}(H_n^{(j)}Z) \le 0, n = 0, \dots, N,\\
& Z_{d+1,d+1} = 1, Z \succeq 0.
\end{aligned} \tag{35}$$

from below.

Lemma 1 *The dual problem of SDP* (35) *takes the form*

$$\max -\gamma, \text{ s.t. } S = \tilde{A}_j(\Lambda) + \tilde{B}_j + \sum_{n=0}^{N} r_n H_n^{(j)} + \gamma e_{d+1}e_{d+1}^T, r \ge 0, S \succeq 0, \tag{36}$$

in variables $r = \begin{bmatrix} r_0 \\ \vdots \\ r_N \end{bmatrix} \in \mathbb{R}^{N+1}$ *and* $\gamma \in \mathbb{R}$.

PROOF Consider the Lagrangian

$$L(Z, r, \gamma) = \operatorname{tr}((\tilde{A}_j(y) + \tilde{B}_j)Z) + \sum_{n=0}^{N} r_n \operatorname{tr}(H_n^{(j)}Z) + \gamma(\operatorname{tr}(Ze_{d+1}e_{d+1}^T) - 1), \tag{37}$$

where $r \in \mathbb{R}_+^{N+1}$ and $\gamma \in \mathbb{R}$. By minimizing $L(Z, r, \gamma)$ w.r.t. $Z \in \mathbb{S}_+^{d+1}$, we derive the dual problem (36).

The constraints on Λ in the dual problem (14) include that the optimal value of (35) is bounded from below by $-\tilde{\beta}$. According to Lemma 1, this constraint is equivalent to that there exist $r \in \mathbb{R}^{N+1}$ and γ such that

$$-\gamma \ge -\tilde{\beta}, S = \tilde{A}_j(\Lambda) + \tilde{B}_j + \sum_{n=0}^{N} r_n H_n^{(j)} + \gamma e_{d+1}e_{d+1}^T, r \ge 0, S \succeq 0. \tag{38}$$

As $e_{d+1}e_{d+1}^T$ is positive semi-definite, the above condition on Λ is also equivalent to that there exist $r \in \mathbb{R}^{N+1}$ such that

$$\tilde{A}_j(\Lambda) + \tilde{B}_j + \sum_{n=0}^{N} r_n H_n^{(j)} + \tilde{\beta} e_{d+1}e_{d+1}^T \succeq 0, r \geq 0. \tag{39}$$

Therefore, the following convex set of Λ

$$\left\{\Lambda : \tilde{A}_j(\Lambda) + \tilde{B}_j + \sum_{n=0}^{N} r_n^{(j,-)} H_n^{(j)} + \tilde{\beta} e_{d+1}e_{d+1}^T \succeq 0,\ r^{(j,-)} \geq 0\right\} \tag{40}$$

is a subset of the set of Λ satisfying the dual constraints

$$\left\{\Lambda : \min_{\|w\|_2 \leq 1, (2D_j - I)w \geq 0} w^T \left(B_j + A_j(\Lambda)\right) w \geq -\tilde{\beta}\right\} \tag{41}$$

On the other hand, the constraint on Λ

$$\max_{\|w\|_2 \leq 1, (2D_j - I)w \geq 0} w^T \left(B_j + A_j(\Lambda)\right) w \leq \tilde{\beta} \tag{42}$$

is equivalent to

$$\min_{\|w\|_2 \leq 1, (2D_j - I)w \geq 0} -w^T \left(B_j + A_j(\Lambda)\right) w \geq -\tilde{\beta}. \tag{43}$$

By applying the previous analysis on the above trust region problem, the following convex set of Λ

$$\left\{\Lambda : -\tilde{A}_j(\Lambda) - \tilde{B}_j + \sum_{n=0}^{N} r_n^{(j,+)} H_n^{(j)} + \tilde{\beta} e_{d+1}e_{d+1}^T \succeq 0,\ r^{(j,+)} \geq 0\right\} \tag{44}$$

is a subset of the set of Λ satisfying the dual constraints

$$\left\{\Lambda : \max_{\|w\|_2 \leq 1, (2D_j - I)w \geq 0} w^T \left(B_j + A_j(\Lambda)\right) w \leq \tilde{\beta}\right\}. \tag{45}$$

Therefore, replacing the dual constraint $\max_{w:\|w\|_2 \leq 1} \left|\sum_{n=1}^{N} \|w\|_2^2 \psi''(w^T x_n) - y_n^T w \psi'(x_n^T w)\right| \leq \tilde{\beta}$ by

$$\begin{aligned}
&\tilde{A}_j(\Lambda) + \tilde{B}_j + \sum_{n=0}^{N} r_n^{(j,-)} H_n^{(j)} + \tilde{\beta} e_{d+1}e_{d+1}^T \succeq 0, j \in [p],\\
&-\tilde{A}_j(\Lambda) - \tilde{B}_j + \sum_{n=0}^{N} r_n^{(j,+)} H_n^{(j)} + \tilde{\beta} e_{d+1}e_{d+1}^T \succeq 0, j \in [p],\\
&r^{(j,-)} \geq 0, r^{(j,+)} \geq 0, j \in [p].
\end{aligned} \tag{46}$$

we obtain the relaxed dual problem. As its feasible domain is a subset of the feasible domain of the dual problem, the optimal value of the relaxed dual problem gives a lower bound for the optimal value of the dual problem.

D.5 Proof of Proposition 5

PROOF Consider the Lagrangian function

$$\begin{aligned}
L(\Lambda, \mathbf{r}, \mathbf{S}) = &-\frac{1}{2}\|\Lambda + Y\|_2^2 - \sum_{j=1}^{p} \operatorname{tr}\left(S^{(j,-)}\left(\tilde{A}_j(\Lambda) + \tilde{B}_j + \sum_{n=0}^{N} r_n^{(j,-)} H_n^{(j)} + \frac{\tilde{\beta}}{2} e_{d+1}e_{d+1}^T\right)\right)\\
&- \sum_{j=1}^{p} \operatorname{tr}\left(S^{(j,+)}\left(-\tilde{A}_j(\Lambda) - \tilde{B}_j + \sum_{n=0}^{N} r_n^{(j,+)} H_n^{(j)} + \frac{\tilde{\beta}}{2} e_{d+1}e_{d+1}^T\right)\right),
\end{aligned} \tag{47}$$

where we write

$$\begin{aligned}
\mathbf{r} &= \left(r^{(1,-)}, \dots, r^{(p,-)}, r^{(1,+)}, \dots, r^{(p,+)}\right) \in \left(\mathbb{R}^{N+1}\right)^{2p},\\
\mathbf{S} &= \left(S^{(1,-)}, \dots, S^{(p,-)}, S^{(1,+)}, \dots, S^{(p,+)}\right) \in \left(\mathbb{S}_+^{d+1}\right)^{2p}.
\end{aligned} \tag{48}$$

Here we write $\mathbb{S}_+^{d+1} = \{S \in \mathbb{S}^{d+1} | S \succeq 0\}$. By maximizing w.r.t. Λ and $\mathbf{r}$, we derive the bi-dual problem (17).

D.6 Proof of Theorem 1

Suppose that (Z, W, α) is a feasible solution to (12). Let $D_{j_1}, \dots, D_{j_k}$ be the enumeration of $\{\mathbf{diag}(\mathbb{I}(Xw_i \geq 0))|i \in [m]\}$. For $i \in [k]$, we let

$$S^{(j_i,+)} = \sum_{l:\alpha_l \geq 0, \mathbf{diag}(\mathbb{I}(Xw_l \geq 0)) = D_{j_i}} \alpha_l \begin{bmatrix} w_l w_l^T & w_l \\ w_l^T & 1 \end{bmatrix}, S^{(j_i,-)} = 0, \tag{49}$$

and

$$S^{(j_i,+)} = 0, S^{(j_i,-)} = -\sum_{l:\alpha_l < 0, \mathbf{diag}(\mathbb{I}(Xw_l \geq 0)) = D_{j_i}} \alpha_l \begin{bmatrix} w_l w_l^T & w_l \\ w_l^T & 1 \end{bmatrix}. \tag{50}$$

For $j \notin \{j_1, \dots, j_k\}$, we simply set $S^{(j,+)} = 0, S^{(j,-)} = 0$. As $\|w_i\|_2 \leq 1$ and $D_{j_i} = \mathbb{I}(Xw_i \geq 0)$, we can verify that $\operatorname{tr}(S^{(j,-)} H_n^{(j)}) \leq 0, \operatorname{tr}(S^{(j,+)} H_n^{(j)}) \leq 0$ are satisfied for $j = j_1, \dots, j_m$ and $n = 0, 1, \dots, N$. This is because for $n = 0$, as $H_0^{(j_i)} = \begin{bmatrix} I_d & 0 \\ 0 & -1 \end{bmatrix}$, it follows that

$$\begin{aligned} \operatorname{tr}(S^{(j_i,+)} H_0^{(j_i)}) &= \sum_{l:\alpha_l \geq 0, \mathbf{diag}(\mathbb{I}(Xw_l \geq 0)) = D_{j_i}} \alpha_l(\|w_l\|^2 - 1) \leq 0, \\ \operatorname{tr}(S^{(j_i,-)} H_0^{(j_i)}) &= -\sum_{l:\alpha_l < 0, \mathbf{diag}(\mathbb{I}(Xw_l \geq 0)) = D_{j_i}} \alpha_l(\|w_l\|^2 - 1) \leq 0. \end{aligned} \tag{51}$$

For $n = 1, \dots, N$, we have

$$\begin{aligned} \operatorname{tr}(S^{(j_i,+)} H_0^{(j_i)}) &= \sum_{l:\alpha_l \geq 0, \mathbf{diag}(\mathbb{I}(Xw_l \geq 0)) = D_{j_i}} 2\alpha_l(1 - 2(D_{j_i})_{nn}) x_n^T w_l \leq 0, \\ \operatorname{tr}(S^{(j_i,-)} H_0^{(j_i)}) &= -\sum_{l:\alpha_l < 0, \mathbf{diag}(\mathbb{I}(Xw_l \geq 0)) = D_{j_i}} \alpha_l(1 - 2(D_{j_i})_{nn}) x_n^T w_l \leq 0. \end{aligned} \tag{52}$$

Based on the above transformation, we can rewrite the bidual problem in the form of the primal problem (13). For $S \in \mathbb{S}^{d+1}$, we note that

$$\begin{aligned} &\operatorname{tr}(S\tilde{A}_j(\Lambda)) \\ =& -\operatorname{tr}((\Lambda^T D_j X + X^T D_j \Lambda) S_{1:d,1:d}) \\ =& -2\operatorname{tr}(\Lambda^T D_j X S_{1:d,1:d}), \end{aligned}$$

where $S_{1:d,1:d}$ denotes the $d \times d$ block of S consisting the first d rows and columns. This implies that $\tilde{A}_j^*(S) = -2D_j X S_{1:d,1:d}$. Hence, we have

$$\tilde{A}_{j_i}(S^{(j_i,+)} - S^{(j_i,-)}) = -\sum_{l:\mathbf{diag}(\mathbb{I}(Xw_l \geq 0)} 2\alpha_l D_{j_i} X w_l w_l^T = -\sum_{l:\mathbf{diag}(\mathbb{I}(Xw_l \geq 0)} 2\alpha_l (Xw_l)_+ w_l^T.$$

Therefore, we have

$$\sum_{j=1}^p \tilde{A}_j^*(S^{(j,-)} - S^{(j,+)}) = 2\sum_{i=1}^m \alpha_i (Xw_i)_+ w_i^T.$$

As n-th row of Z satisfies that $z_n = 2\sum_{i=1}^m \alpha_i w_i (x_n^T w_i)_+$, this implies that

$$Z = 2\sum_{i=1}^m \alpha_i (Xw_i)_+ w_i^T = \sum_{j=1}^p \tilde{A}_j^*(S^{(j,-)} - S^{(j,+)}).$$

Hence $(Z, \{(S^{(j,-)}, (S^{(j,-)}\}_{j=1}^p)$ is feasible to the relaxed bi-dual problem (17).

We can also compute that

$$\sum_{j=1}^p \operatorname{tr}(\tilde{B}_j(S^{(j,+)} - S^{(j,-)})) = 2\sum_{i=1}^m \alpha_i \sum_{n=1}^N \mathbb{I}(x_n^T w_i \geq 0)\|w_i\|_2^2,$$

and

$$\sum_{j=1}^{p} \operatorname{tr}\left((S^{(j,+)} + S^{(j,-)}) e_{d+1} e_{d+1}^{T} \right) = \sum_{i=1}^{m} |\alpha_i|.$$

Thus, the primal problem (13) with (Z, W, α) and the relaxed bi-dual problem (17) with $(Z, \{(S^{(j,-)}, (S^{(j,-)}\}_{j=1}^{p})$ have the same objective value.